Acta Linguistica Asiatica

Year 2014, Volume 4, Number 2

ACTA LINGUISTICA ASIATICA
Year 2014, Volume 4, Number 2

Published by: Znanstvena založba Filozofske fakultete Univerze v Ljubljani
(Ljubljana University Press, Faculty of Arts)

Issued by: Department of Asian and African Studies

For the publisher: Branka Kalenić Ramšak, the Dean of the Faculty of Arts

Journal's web page:
http://revije.ff.uni-lj.si/ala/
The journal is published in the scope of Open Journal Systems

ISSN: 2232-3317

Abstracting and Indexing Services:
COBISS, dLib, Directory of Open Access Journals, MLA International Bibliography, Open J-Gate and Google Scholar.

Publication is free of charge.

Address:
University of Ljubljana, Faculty of Arts
Department of Asian and African Studies
Aškerčeva 2, SI-1000 Ljubljana, Slovenia

E-mail: nina.golob@ff.uni-lj.si

TABLE OF CONTENTS

FOREWORD

This issue is dedicated to the field of corpus linguistcs, a relatively young but well-eastablished approcah to the study of language, and particularly to researches that base on the Japanese BCCWJ corpus.

The Balanced Corpus of Contemporary Written Japanese (BCCWJ) – a one hundred million word corpus released by the National Institute of Japanese Language and Linguistics (NINJAL) in the year 2011 – has greatly advanced research in the field of Japanese language. Utilizing a corpus that is made-up of a multitude of actual language data enables researches to overcome the limitation of basing their analysis primarily on either intuition or scantly language data, making it possible to study languages in a more objective and comprehensive manner. In addition to this, analyzing data compiled by collecting learner essays or speech makes it possible to throw light on errors based on mother-tongue influence as well as acquisition patterns independent of mother-tongue influence. Furthermore, the observations and analysis based on actual language use can be used to create dictionaries and other authentic study materials.

Though the compilation of a Japanese language corpus falls far behind when compared to corpora in English, a fairly large number of Japanese corpora like the BCCWJ have been made available in recent years. Also, a corpus of ten billion words is currently being complied at NINJAL making it possible to utilize a large-scale corpus of greater diversity in the future. Such corpora may be utilized in the following research field pertaining to Japanese language:

- Japanese language research
- Analyzing Japanese language syllabus
- Determining difficultly levels of words and passages
- Developing tools that aid text-comprehension and essay writing
- Developing tools that aid corpus-driven learning
- Automatic assessment and correction of learner errors

The thematic issue thus introduces four research studies that utilize the above mentioned corpus in the field of Japanese language education. The final research is a good example of a research of languages in contact and is also based on a corpus, though a non-electronic one.

Andrej Bekeš discusses spatio-temporal metalinguistic labels, NSM (natural semantic metalanguage), introduces a methodology based on the corpus-based approach that emphasizes function and meaning and the approach that emphasizes relation to

Acta Linguistica Asiatica, 4(2), 2014.
ISSN: 2232-3317, http://revije.ff.uni-lj.si/ala/
DOI: 10.4312/ala.4.2.5-6

context, and presents his observations on metalinguistic terms for dictionaries of Japanese expressions.

Irena Srdanović introduces various tools that search and retrieve collocations and presents her analysis on collocations that utilizes Japanese corpora and corpus tools by giving actual examples of collocations consisting of adjective and noun combinations.

Kikuko Nishina and her colleagues present an overview of the three systems developed through the Hinoki Project at the Tokyo Institute of Technology, namely Natsume (collocation search system), Natane (a learner corpus search system that supports Natsume) and Nutmeg (an automatic correction, composition support system), while emphasizing the need for learner corpus.

Kristina Hmeljak Sangawa introduces exercises developed by utilizing Japanese monolingual corpora and Japanese-Slovenian parallel corpus in order to deepen the understanding of vocabulary meaning, its structure and style.

Anrei A. Avram attempts to make an overall description of the phonological, morphological, syntactic features, and vocabulary in 19th century Yokohama Pidgin Japanese by using written corpora.

Yuriko Sunakawa

RESEARCH ARTICLES

RETHINKING METALINGUISTIC LABELS: SPATIO-TEMPORAL METALINGUISTIC TERMS IN LEARNERS' DICTIONARIES OF JAPANESE EXPRESSIONS

Andrej BEKEŠ
University of Ljubljana
andrej.bekes@gmail.com

Abstract

Paper examines spatio-temporal metalinguistic terms in learners' dictionaries of Japanese expressions in major existing dictionaries. Based on the analysis it proposes a layered metalinguistic labelling solution to achieve the greates efficiency with the smallest possible number of labels being employed.

Keywords: Japanese; metalinguistic label; semantic/functional index; grammar/ expression dictionaries

Povzetek

Članek analizira prostorsko-časovne metajezikovne izraze v slovarjih japonskih izrazov za učeče se. Na osnovi analize predlaga večplastno uporabo metajezikovnih oznak, s čemer se optimizira učinkovitost ob kar se da majhnem naboru oznak.

Ključne besede: japonski jezik; metajezikovna oznaka; semantična/ funkcijska razvrstitev; slovarji slovnice/ izrazov

1. Introduction

Efficient semantic labelling is essential if a dictionary is to be used for the composition as well as the reception of texts. When creating a learners' dictionary, this task tends to be done based on experience and tradition, and often in quite an ad hoc fashion, that is without systematic theoretical considerations (e.g. Group JAMASSY 1998 where I was in charge of semantic/functional labels).

One option for a more systematic treatment is offered by the NSM (natural semantic metalanguage) approach (e.g. Goddard and Wierzbicka 2007, for a critical assessment c.f. Trobevšek Drobnak 2009). Another possibility is to build an array of metalinguistic labels from the bottom up, from corpora (e.g. Labrador De La Cruz 2004). On the other hand, the chosen labels should also be as theoretically relevant as possible.

Acta Linguistica Asiatica, 4(2), 2014.
ISSN: 2232-3317, http://revije.ff.uni-lj.si/ala/
DOI: 10.4312/ala.4.2.9-23

This means that if adequate labels are to be produced, not only collocations but also broader semantic properties, implicit within the wider context, need to be considered.

Languaging (e.g. Becker 1988) is an activity where our linguistic potential is applied to fluid on-going situations. Not only functional expressions but also lexical items such as verbs, adjectives and nouns, often have multiple meanings, which arise through interplay with the context. The problem is how to capture these characteristics in a metalanguage intended for explanation of the lexicon, or when providing learners with semantic/functional labels.

One of the drawbacks of NSM is its apparently static view of the lexicon and the meanings of lexical items, missing a crucial property of human communication. Regarding methodology, the intuitively chosen approach of Group JAMASSY (1998) to provide meanings not so much by description as by illustration, giving many examples of particular uses of functional lexical items, seems appropriate. While more terse than Makino and Tsutsui (2008) in the sense that its explicit explanations are limited to short commentaries, it seems to be a less restrictive approach (than what?) when guiding experienced users(of what?)/learners.

The context-based approach, resonating with the basic argument in Labrador De La Cruz (2004), has recently been taken up systematically. Limiting the discussion just to the field of Japanese language learning support tools, two outstanding such tools should be noted. One is Natsume (*http://hinoki.ryu.titech.ac.jp/natsume/*), developed by Kikuko Nishina and associates and running partly on the BCCWJ and partly on closed corpora, and the other is NINJAL LWP (Lago Word Profiler), developed by Prashant Pardeshi and associates and implemented using the BCCWJ and TWC corpora (*http://corpus.tsukuba.ac.jp/search/*). These tools provide users with usage based information, extracted online by means of an analysis of the chosen corpora. For this reason, these tools are actually used not only by intermediate and advanced learners, but also by professional dictionary makers.

In spite of the power of these new tools, well thought-out dictionaries for beginner and lower intermediate learners are also crucial to speed up efficient learning of Japanese. For this purpose, an efficient array of semantic/functional labels that guides the users towards an effective understanding of the lexical item is necessary. The labels should be chosen so that for composition, they can efficiently guide the learner towards those expressions in the target language that are most appropriate to achieve the learner's goal in a given situation. Since learners are not translators, however, this goal should be achieved through learners' intuitive grasp "from within" of how the target language works.

This paper focuses on spatio-temporal metalinguistic terms, with an emphasis on methodological issues, and analyses the implementation of such metalinguistic terms in contemporary Japanese grammar/expression dictionaries, in particular Group JAMASSY (1998). Building on the analysis of how spatio-temporal terms are structured

in Group JAMASSY (1998) I argue for a concise but systematic and pragmatically relevant approach in semantic labelling.

2. Japanese grammar / expression dictionaries

A number of dictionaries (Group JAMASSY, Makino and Tsutsui, Tomomatsu et al. etc.) have been published since 1990 with the express purpose of easing the learning of items that do not appear in traditional dictionaries, i.e., functional expressions, sentence patterns, etc. In this paper, two such dictionaries were chosen, based on their breadth of coverage and popularity. An additional reason is that I am a co-author of one (Group JAMASSY 1998) and this paper serves as a re-examination of that previous work.

2.1 Characteristics of Makino and Tsutsui (2008) and Group JAMASSY (1998)

The purpose of the dictionaries is largely similar: providing clear information to learners about common expressions and patterns.

Makino and Tsutsui authored three volumes, from elementary to advanced, while Group JAMASSY's dictionary targets upper intermediate and advanced level learners. Therefore, only Makino and Tsutsui's third volume (2008), targeting advanced level learners, will be taken into consideration here.

In both dictionaries the main entries include lexical items and constructions, and are organised alphabetically (Makino and Tsutsui) or in *gojūon* order (Group JAMASSY).

Makino and Tsutsui offer detailed information on formation, i.e., the constructions in which particular expressions are typically used. Group JAMASSY handles this type of information in subentries that are often detailed and hence sometimes difficult to follow.

There is one crucial difference in the organisation of the dictionaries. Makino and Tsutsui is bilingual - with Japanese entries and examples but strictly English explanations and glosses. On the other hand, Group JAMASSY's dictionary was conceived as a monolingual dictionary, aimed primarily at upper intermediate and advanced learners and as a reference book for Japanese language teachers. In the wake of its success, translations of the dictionary into Chinese, Korean, Thai and Vietnamese have appeared.

While Makino and Tsutsui provide ample examples of use together with translations and detailed explanations, Group JAMASSY relies on the power of multiple examples, supplemented with direct, concise explanations. In the translated versions of the dictionary the examples and the explanations are translated into the respective language, resulting in a dictionary of a structure similar to Makino and Tsutsui.

Another difference is in the organisation of the index. Makino and Tsutsui (2008), being bilingual, offers two indexes, an index of romanised Japanese entries in alphabetical order, and an index of English translations of entries, also in alphabetical order. This solution may seem helpful for users engaging in productive tasks, but is not always conducive to a quick location of the desired expression in Japanese.

Group JAMASSY tries to cope with the problem of quick entry location by including two indexes: besides the *gojūon* index of entries, there is also an index of entries sorted by their meaning and/or function. This semantic/functional index consists of more than one hundred metalinguistic labels, referring broadly to the meaning and/or function of a particular entry or subentry. In this respect, Group JAMASSY (1998) is similar to some of the other grammar dictionaries that appeared later, such as Tomomatsu et al. (2007/2010).

Below, this paper focuses on the subset of these labels that refer to spatial concepts, temporal concepts, or both, in Group JAMASSY (1998).

3. Metalinguistic labels

Attempts at creating a language that would precisely convey meaning across languages have a distinguished lineage, going back to the tradition of searching for a universal language - with G. W. Leibnitz to be mentioned as one prominent proponent of the idea (e.g. Yaguello 1990, Eco 2003), and with NSM (e.g. Goddard and Wierzbicka 2007) being a notable present attempt.

In both dictionaries under scrutiny, the explanations should have ideally been based on some subset of language in the vein of NSM, but this is not he case with Group JAMASSY and does not seem to be the case with Makino and Tsutsui either. In both dictionaries ordinary, though to some extent controlled, language is used for the explanations.

3.2 The purpose of the metalinguistic labels in Group JAMASSY dictionary

Group JAMASSY (1998) uses metalinguistic labels to hint at the meaning and/or function of entries and subentries. There are two types of labels: (a) labels used in the semantic/functional index, to direct the learner towards the relevant entries during language production tasks, and (b) those used in the main body of the dictionary, to disambiguate different meanings and usages.

The problem is that both sets of labels are not completely identical and are used in a way that seems somewhat inconsistent. Let us consider for example the entry あいだ.

(1) a <Index labels>
期間

b <Entry labels>

【あいだ】
1 N のあいだ
a N のあいだ ＜空間＞
.....
b N のあいだ ＜関係＞ <most frequent use in TWC>
.....
2 あいだ ＜時間＞
a ...あいだ
[N のあいだ]
[A-いあいだ]
[V-ている/V-るあいだ]

In the semantic/functional index the entry あいだ is listed under the label ＜期間＞ (meaning interval of time, period, see (1)a above), while in the main body, あいだ is disambiguated with the synonymous / hypernymous ＜時間＞.

In the index one also finds labels such as ＜無関係＞, ＜前後関係＞ and ＜空間的関係＞, while in the body of the dictionary, as in example in (1)b, the hypernym ＜関係＞is used, which in this context refers to the specific meaning of a “human relationship” and happens to be the most frequent use in TWC.

Also, while for example ＜空間＞ and ＜時間＞, as above, are often used in the main body of the dictionary, there are no such labels in the index. Here, on the other hand, one can find ＜時点＞, ＜期間＞, ＜空間的関係＞, which point to the relevant entries in the body.

In this way, there are 158 labels used only in the main body, 115 labels used only in the index, and 45 labels that are used both in the index and in the main body of the dictionary.

3.3 Clashing requirements

There is a contradiction inherent in devising metalinguistic expressions to be used for explanation in dictionaries. On the one hand, the number of such expressions has to be small, so that they are manageable, and on the other hand, for these expressions to be specific enough, their number has to be sufficiently large. This is the clash of requirements that Leibnitz and others had to face early in their work on universal language (e.g. Eco 2003; Yaguello 1990). The problem that for a while also hindered attempts at automatic translation is that if one wants to make a language with the expressive power of natural languages it has to be as complex as natural languages.

Group JAMASSY's attempt was no exception. Its goal was simplicity and intuitive ease of use, which resulted in plentiful examples, concise explanations, and a very limited set of non-technical transparent semantic/functional labels used in the main body and index of the dictionary. The request that labels in the index should be specific enough to direct the learner to a small set of potential candidate entries is in conflict with the requirement that the number of labels should be kept at a manageably small number. The outcome of the clash was the aforementioned inconsistency between the labels used in the index and those used in the body of the dictionary.

Though this is theoretically an unresolvable problem (NSM being no exception), a trade-off is possible. The compromise would be to establish an array of labels sufficiently rich for practical purposes but at the same time also concise enough to be manageable.

Randomly chosen examples such as (1) above show that an improvement in the consistency and accuracy of semantic/functional labels is possible.

4. Some proposals for improvement

Limiting the scope to spatio-temporal labels, I will illustrate a few possible paths out of this impasse.

4.4 Consistency of use - labels in the body and in the index

There are many examples where the use of semantic/functional labels could be made more consistent by treating the labels in the body of the dictionary and the labels in the index as parts of the same set. In the following sections, examples based on the use of spatio-temporal labels are presented.

4.4.1 Unifying labels in the dictionary body and in the index

The labels in the index and the labels used for disambiguation in the body of the dictionary are sometimes related (hypernym-hyponym etc...), yet differ. For example, there are cases when 期間 is used in the index while 時間 is used in the body of the dictionary. The first entry in the dictionary is a good example:

(2) 2 あいだ <時間> p. 2
a ...あいだ
[N のあいだ]
[A-いあいだ]
[V-ている/V-るあいだ]
b ...あいだに
[N のあいだに]

[Na なあいだに]
[A-いあいだに]
[V-ている/v-るあいだに]

In (2) 2a, the meaning of あいだ can be glossed as "an interval of time during which an action continues" (継続動作が続く/いた期間). It would seem to be more user friendly to use the same label for disambiguation and in the index, and 期間 is more specific than 時間. In 2b, the meaning of あいだに can be glossed as "(relatively) instantaneous action, taking place within an interval of time" (瞬間動作が起る/った期間), so again, as in 2a, it seems appropriate for the labels to be the same, i.e., 期間. A more or less identical situation seems to be the case in (3) below, semantically similar to the preceding example, with the meaning of うちに glossed as "(relatively) instantaneous action, taking place within an interval of time" (瞬間動作が起る/った期間):

(3) 2 うち <時間> p. 48
[N のうちに]
[Na なうちに]
[A-いうちに]

In example (4) and (5) below, the situation is similar again. The meaning of both constructions can be glossed as "an interval of time during which an action or state continues" 状態・継続動作が続く/いた期間. In (4), as in (2)2a, using 期間 for disambiguation would be more appropriate.

(4) 2 N ごしの <時間> p. 110
(5) 2 N じゅう <時間> p. 145

On the other hand, in (6), both keeping and changing the disambiguation label could be argued for:

(6) 3...まで <時間> p. 546
a N まで
b V-るまで

The gloss of まで in 3a in (6) is “the time until which a state or some activity continues” (状態・継続動作の時限), and the gloss of まで in 3b is “the time limit defined by an action, until which a state or some activity continues” (態・継続動作の、別の動作によって限定される時限). In both cases, the hypernym 時間 seems acceptable from the point of view of label economy, though a more specific label, such as “time limit” 時限 would be even more appropriate. Additionally, in 3b, the temporal-relational semantic relationship between the defining action and continuing state or activity could also be emphasized, by means of double labelling, a possibility discussed in more detail in Section 3.2.

Similarly, the construction in the example below:

(7) 2 N を N にひかえて <場所> p. 505

expresses the general meaning of a spatial relationship between a location and its surroundings. Therefore the more specific "place, location" <場所> could be replaced by <空間> as a disambiguator.

4.4.2 Consistent disambiguation of subentries

Consistent disambiguation of the subentries would make searching easier and improve understanding of the entries. The first entry in the dictionary is once again a good example.

(8) 【あいだ】
1 N のあいだ p. 2
a N のあいだ <空間>
b N のあいだ <関係>
2 あいだ <時間>
a .....あいだ
b ...あいだに
...

In (8) the macrostructure of the entry 【あいだ】 is shown. While the temporal use of あいだ is disambiguated at the first level of the subentry, the spatial and relational use are disambiguated at the second level of the subentries. Restructuring the entry 【あいだ】 into

(8') 【あいだ】
1 N のあいだ <空間>
2 N のあいだ <関係>
3 あいだ <時間>
a ...あいだ
b ...あいだに
...

would render the treatment of subentries in the dictionary more consistent.

Restructuring the entry would also be advisable in cases such as 【まで】 below:

(9) 【まで】 p. 545
1 N から N まで <Ø>
(1) シンポジウムは 1時から 3時まで第 3会場で行います。
(2) A :大阪から東京までどのくらいかかりますか。

.....
2 N まで <目的地>
3...まで <時間> p. 546
a N まで
b V-るまで
4...まで <程度> p. 547

In (9) the first level subentry 1 is not disambiguated at all, while giving examples of the temporal and spatial use of N から N まで. Such situations may discourage an inexperienced user, who otherwise could make good use of such a dictionary, if its entries were organised in a more consistent and transparent way. A more appropriate entry structure would be:

(9') 【まで】
1 N から N まで <時間>
(1) シンポジウムは 1 時から 3 時まで第 3 会場で行います。
.....
2 N から N まで <空間>
(2) A :大阪から東京までどのくらいかかりますか。...
.....
3 N まで <目的地>
4...まで <時間>
a N まで
b V-るまで
5..まで <程度>
...

4.5 Combination of labels for greater precision

One way to increase the expressive power of a limited number of semantic/functional labels is to combine the labels. For example, a superordinate label could specify a wider semantic area, while the subordinate label of the same entry would narrow the focus to one of its aspects. Incongruent examples, using a partly overlapping set of labels in the index and in the main body, such as (10) below, provide a hint:

(10) index: 条件(確定条件); 条件(仮定条件); 条件(一般条件)
【とすると】 p. 34
main body:
1 ... とすると <仮定条件>
2 ... とすると <確定条件>
.....

The index entry, which at present looks like:

(11) 条件(一般条件) と
.....

条件(仮定条件)　　かりに /...たら/...ば
.....
条件(確定条件)　　とすると

would look like:

(11') 条件
〈一般条件〉　　と
...
〈仮定条件〉　　かりに /...たら/...ば
...
〈確定条件〉　　とすると

while the labelling in the main body of the dictionary would not change. With a two-layered labelling in the index the user would have a clearer idea of how entries are interrelated.

Below I offer some additional examples of temporal and spatial labels that could become more transparent by being rearranged into two layers.

In the index entry (12) below, the spatial use of the entry expression ごし is indicated by the label 空間的関係, while there is no label provided to indicate the temporal use of ごし.

(12) 【ごし】　　p. 110
1 N ごし<空間>
2 N ごし<時間>

In the "spatial" use of ごし the relational meaning seems to be more general, and it should take precedence over the spatial meaning. As for the temporal use of ごし, as shown in example (4) above, it implies an interval. The temporal meaning should therefore be labelled as 期間. The index should hence be rewritten as follows:

(12') 関係
〈空間的関係〉
ごし
...
期間
ごし
...

The main entry would then look like:

(12”) 【ごし】
1 N ごし <空間的関係> → <N(物）を間において>の行為
2 N ごし <期間> → 状態・継続動作が続く/いた期間

There is no end to possible examples. The next example is again an entry with both spatial and temporal meanings.

(13) 【じゅう】 p. 145
1 N じゅう（に） <空間>
2 N じゅう <時間>

In the index, the entry じゅう is referred to by the labels 空間的関係 and 期間. In its spatial use, the meaning seems not to be so much relational as implying “the range or scope of some activity” (動作が及ぶ範囲、空間), thus 範囲 is appropriate, subspecified for space <空間>. The temporal use label, <時間>, seems too vague for the gloss “the period while a situation or an action continues” (状態・継続動作が続く/いた期間). Both the spatial and temporal uses are related; 期間 is a sort of temporal range, that is, a 範囲. The revised index labelling and entry should thus look as follows:

(13’) index:
範囲
<空間>
じゅう;
期間
じゅう
entry:
【じゅう】
1 N じゅう（に） <範囲-空間>
2 N じゅう <期間> → 状態・継続動作が続く/いた期間

In this section, a two-layered labelling was proposed. As can be seen from example (14), finer subdivisions with more layers of labels are possible and sometimes necessary. In (14) the entry 【あと $_1$】 is referred to in the index by two labels, 空間的関係 and 前後関係. The problem is that 前後関係 can refer to both temporal and spatial order, so as a label it is not very useful for the learner.

(14) 【あと $_1$】 p. 8, 9
1 あと <空間>
[N のあと]
[V-る/V-たあと]
2 あと(で/に) <時間>
a ...あと
[N のあと] 動作(N)の継起

[V-たあと] 動作(V)の継起
...

Two-layered labelling would provide clearer information for the learner. The entry あと 1 is about temporal and spatial order, so rather than 時間 or 空間, its relational property should be given priority in the choice of labels. In the index, the labels should be arranged as below, in (14'). For greater precision and a sharper focus, an additional layer can be added, in this case, <前後関係>. This is the reverse of the original index, where the ambiguous label <前後関係> appears as the first label. Thus we would have in the index:

(14') 関係
〈空間的関係〉
〈前後関係〉
あと
〈時間的関係〉
〈前後関係〉
あと

and in the body:

(14") 【あと 1】
1 あと <空間的関係-前後関係>
[N のあと]
[V-る/V-たあと]
2 あと(で/に) <時間的関係-前後関係>
a ...あと
[N のあと]
[V-たあと]

4.6 Contribution of corpora

Frequency data from corpora can be a guide for ordering different usages, though this principle is not an absolute guide. Let us consider entry (1) once again in (15).

(15) 【あいだ】
1 N のあいだ
a N のあいだ ＜空間＞
.....
b N のあいだ ＜関係＞
2 あいだ ＜時間＞
a ...あいだ
.....

Here for example frequency based ordering and semantic (more general first) ordering clash. Data from Lago TWC show that the use as RELATION is more frequent than the use as SPACE. On the other hand, semantically, SPACE is primary and RELATION is derived.

Further, the semantic properties of N in N のあいだ may also be relevant. The choices in the Group JAMASSY do not always seem to be the most appropriate. Here are some representative examples, based on the same entry:

(16) b N のあいだ〈関係〉 p.2 <Group JAMASSY>
(1) 最近二人の間はうまくいっていないようだ。
(2) そのホテルは安くて清潔なので、旅行者たちの間で人気がある。
(3) 二つの事件の間にはなにか関係があるらしい。

Considering the first example above, (16)(1), frequencies in LAGO TWC are 402 for N=二人 and 1535 for N=人々. This difference should be big enough to warrant the choice of 人々instead as the typical example of relation involving members of a set. On the other hand, the second example of use (16)(2) seems to be appropriate as far as the choice of N is concerned, as Ns that belong to the semantic category HUMAN (＜人間＞) represent the vast majority of examples of use in LAGO TWC. There is doubt, however, surrounding whether (16)(2) is indeed an example of RELATION ＜関係＞. It seems that could be interpreted rather as expressing a RANGE (or scope)＜範囲＞ where the predication relationship is between「ホテル」 and「人気がある. The third example of use, (16)(3) shows a relation between abstract Ns. The frequency in LAGO TWC for N= **<** abstract noun **>** is very small (< 100), compared to N=人々(1535), but the example seems to be appropriate, because abstract nouns are a separate category.

In accordance with Labrador De La Cruz (2004), labels based on corpora may improve the overall quality of labelling. Abstractions should be based on empirically obtained examples of use and on the goals of a particular dictionary. Some examples of 間 from LAGO TWC follow:

(17)
a 夫婦の間がギクシャク。 <frq. 583>
b 異なる国の間で行う売買取引。 <frq.209>
c この「縁と契約」という二つの文化の間には、天と地ほどの乖離がある。

Here, (17) a is a relation between two specific sentient beings with volition, also quite a frequent pattern observed in the corpus. (17)b is a relation between two possibly not well defined human groupings, but still rather frequent. (17)c is a relation among abstract categories, denoting general characteristics of some human society, and such constructions are not at all frequent in the corpus. From the three examples in (17) it is possible to see that the general label RELATION could be subspecified further as VOLITIONAL (in 17a), as GENERALISED VOLITIONAL (in 17b) and as PERCEIVED/ASCRIBED (in 17c). How far this labelling process should go in practice

is perhaps one of the most difficult tasks of future dictionary making, not only for JAMASSY 2.0.

5. Conclusions

In Section 1 I compared the organisation of Group JAMASSY (1998) with that of Makino and Tsutsui (2008), pointing out the differences stemming from the different basic concepts on which the dictionaries are based. Makino and Tsutsui's volumes are more like grammar handbooks, arranged in dictionary form, with ample examples and explanations. On the other hand Group JAMASSY's dictionary is organised as a dictionary of entries that are not described in general dictionaries. While the two dictionaries overlap, Group JAMASSY covers a larger number of entries. It offers plentiful examples with concise explanations. It also offers relatively systematic semantic/functional labels, used in the index to guide the user towards the desired entries, and within the entries, to disambiguate and explain usage.

In section 2, I briefly examined some issues connected with metalinguistic labels, in particular the inherent limitations of "universal" semantic/functional labels, and clashing requirements that force editors to find a pragmatic balance between accuracy and usability.

Section 3 dealt with the metalinguistic labels in the Group JAMASSY dictionary. Consistency of structuring and use is a precondition for the successful application of labels. Some inconsistencies in the organisation and choice of labels were pointed out and alternatives suggested. Further, having two sets of labels, one for the index and one for the main body of the dictionary, was shown to be inappropriate. As a solution, a merger into one consistently organised set was proposed, such that it could be used to create consistent disambiguation of entries and subentries, while using the same labels also for the meaning and function index. The merger of labels has a drawback, i.e., an increase in the number of labels. To keep the number of labels low and at the same time efficient for accessing the entries from the index via meaning/function, or as transparent disambiguators in the main body of the dictionary, a layered use of labels was proposed and illustrated with examples of spatio-temporal entries. Lastly, some possibilities of how to use corpus data were also outlined.

Literature

Becker, A.L. (1988). Language in particular: A lecture. In D. Tannen (Ed.), *Linguistics in context: connecting observation and understanding*, pp. 17-35. Norwood, NJ: Ablex Publishing Corporation.

Eco, Umberto (2003) *Iskanje popolnega jezika v evropski kulturi* (La ricerca della lingua perfetta nella cultura europea, Bari: Laterza 1993, transl. by Vera Troha). Ljubljana: Založba /*cf.

Goddard, Cliff and Anna Wierzbicka (2007). Semantic primes and cultural scripts in language learning and intercultural communication. In Gary Palmer and Farzad Sharifian (eds.), *Applied Cultural Linguistics: Implications from second language learning and intercultural communication*, pp.105-124. Amsterdam: John Benjamins.

Group JAMASSY (1998) *Nihongo bunkei jiten*. Tokyo: Kurosio.

Makino, Seiichi and Michio Tsutsui (2008). *A Dictionary of Advanced Japanese Grammar*. Tokyo: The Japan Times.

Labrador De La Cruz, Belén (2004). A Methodological Proposal for the Study of Semantic Functions across Languages. *Meta: journal des traducteurs / Meta: Translators' Journal*, Vol. 49/ 2, pp. 360-380.

Tomomatsu Etsuko et al. (2007/2010) Shinsōban donna toki dō tsukau nihongo hyōgen bunkei jiten. Tokyo: ARK.

Trobevšek Drobnak, Frančiška (2009). On the Merits and Shortcomings of Semantic Primes and Natural Semantic Metalanguage in Cross-Cultural Translation. *English Language Overseas Perspectives and Enquiries*, Vol. VI/1-2, pp. 29-41.

Yaguello, Marina (1990) *Gengo no musōsha: 17seiki fuhengengo kara gendai SF made* (Les fous du langage: des langues imaginaires et de leurs inventeurs. Paris: Seuil 1984, transl. by Tanikawa Taeko, Eguchi Osamu). Tokyo: Kosakusha.

CORPUS-BASED COLLOCATION RESEARCH TARGETED AT JAPANESE LANGUAGE LEARNERS

Irena SRDANOVIĆ
University of Ljubljana
irenasrdanovic@gmail.com

Abstract

This paper discusses corpus-based research on collocations, introduces various tools for querying and extracting Japanese collocations and presents an analysis of Japanese collocations using language corpora and related tools. First, major corpus query tools such as Sketch Engine, NINJAL-NLP, Natsume, Chunagon, which can be used by learners and teachers of Japanese language, are briefly described. Focus then shifts to adjectival and nominal collocates and the resource "Collocation data of adjectives and nouns" which consists of adjective headwords and their nominal collocates extracted from two large corpora, BCCWJ and JpTenTen: 500 adjectives and 9,218 collocate nouns, and 500 adjectives and 23,220 collocate nouns from each corpus respectively. Finally, it is shown that corpus-based resources can be used in the creation of reference materials for learners of the Japanese language. The benefits of empirical research into collocations are also shown by comparing the obtained results with collocations in textbooks for Japanese as foreign language.

Keywords: Japanese language; collocations; second-language acquisition; language learning; corpora and tools

Povzetek

Članek razpravlja o raziskavah o kolokacijah, ki temeljijo na korpusnih podatkih, predstavlja različna orodja za poizvedovanje in pridobivanje kolokacij japonskega jezika ter predstavlja analizo kolokacij na podlagi korpusov in orodij za japonski jezik. V članku, bom na kratko opisala najbolj pomembna orodja za iskanje po korpusih, kot so Sketch Engine, NINJAL-NLP, Natsume, Chunagon, ki jih učenci in učitelji japonski jezik lahko uporabljajo. Naprej se osredotočim na kolokacije pridevnikov in samostalnikov in predstavim vir "Kolokacije pridevnikov in samostalnikov", ki je sestavljen iz pridevnikov, kot iztočnic in njihovih samostalniških besednih zvez, pridobljenih iz dveh velikih korpusov BCCWJ in JpTenTen: 500 pridevnikov in 9301 samostalnikov ter 500 pridevnikov in 23.247 samostalnikov, iz obeh korpusov v tem zaporedju. Nazadnje bom pokazala, kako se podatki pridobiti iz korpusa lahko uporabljajo pri ustvarjanju materialov za učence japonskega jezika in kako so lahko koristne empirične raziskave na področju kolokacij, s primerjanjem dobljenih rezultatov z kolokacijah v učbenikih za japonščino kot tuji jezik.

Ključne besede: japonski jezik; usvajanje drugega jezika; učenje jezika; kolokacije; korpusi in orodja

Acta Linguistica Asiatica, 4(2), 2014.
ISSN: 2232-3317, http://revije.ff.uni-lj.si/ala/
DOI: 10.4312/ala.4.2.25-36

1. Introduction

The importance of collocation research has been recognized especially within corpus linguistics with the development of empirical research methods. There are also an increasing number of Japanese language resources and studies on collocations, for example about which collocations are easily acquired and which are burdensome for foreign language learners or about the need to systematically teach collocations to learners and provide reference materials on collocations to ease the process of learning (Oso and Takizawa 2003, Himeno 2012, Srdanović 2013a).

In this paper, I first briefly describe various tools for searching and extracting Japanese collocations in corpora (Chakoshi, Case frame, Sketch Engine, Natsume, NINJAL-NLP, Chunagon). Next, I focus on collocations of the type adjective+noun and present the resource "Collocation data of adjectives and nouns" which consists of adjectival headwords and their noun collocates extracted from two large corpora, BCCWJ and JpTenTen: 500 adjectives and 9,218 collocate nouns, and 500 adjectives and 23,220 collocate nouns from each corpus respectively.

I then will discuss the types of information that can be provided in a hypothetical dictionary of collocations for Japanese language learners, such as placing emphasis on collocations that are difficult for language learners to predict, displaying lexical-map information, corpus-informed information on register and special usages, and so on. Finally, the comparison of these research results with collocations in textbooks for Japanese as foreign language indicates the importance of empirical research into and systematic treatment of collocations, which can then be applied in the creation of corpus-based Japanese language learning materials.

2. Corpus query tools for collocation analyses

Kawahara and Kurohashi (2006) developed the Case Frame [1] online search functionality that initially only extracted verbs and their case frames from the corpus but was later on expanded. Currently it provides a valuable predicate-argument profile of target words from a large web corpus, although the corpus data is limited to extracted sentences only. The first corpus query tool specifically created for Japanese collocation extraction, Chakoshi,[2] was developed at Nagoya University, and is a Japanese text search and collocation extraction system (Fukuda 2007). However, the public version of this tool is limited to searches of the Aozora bunko corpus of literary texts and the Nagoya kaiwa corpus.

[1] Case Frame: http://lotus.kuee.kyoto-u.ac.jp/cf-search/

[2] Chakoshi: http://tell.cla.purdue.edu/chakoshi/public.html

In 2007, the first corpus-query system with detailed lexical profiles of search words for Japanese appeared (Srdanović et al. 2008), Sketch Engine.[3] Sketch Engine was originally created for English (Kilgarriff et al. 2004), but then more than a hundred corpora and sketch grammars for other languages were gradually added to the system. As described by McEnery and Hardie (2012), the tool belongs to the fourth generation of corpus-query tools with various advanced search functionalities. The major functionality breakthrough was called word sketches, "one-page automatic, corpus-based summaries of a word's grammatical and collocational behaviour" (Kilgarriff et al. 2004).[4]

At this stage, Sketch Engine can be searched for automated summaries of Japanese collocational relations within the 400-million word JpWaC (Srdanović et al. 2008) and the two versions of JpTenTen (Srdanović et al. 2013) - the full one, a 10-billion token corpus with short-unit-word annotation, and the smaller sample one, a 200-million token corpus with long-unit-word annotation (suw and luw). This tool enables inclusion and creation of other corpora for personal or group use, so the author got permission to use BCCWJ with Sketch Engine as well. Other functionalities include SketchDiff - comparison of the collocational behavior of two similar words, thesaurus, corpora comparison, parallel corpora concordancing, and lexicographic tickbox (Kilgarriff et al. 2004).

The Japanese word sketches inspired researchers in Japan to create online tools that automatically summarize collocations. Unlike the multilanguage-oriented Sketch Engine, these tools are for Japanese only, but have other specific functionalities and are available for free. These tools, Natsume and NINJAL-LWP, are described below.

Natsume[5] is a computer-assisted language learning system that supports learners' Japanese writing skills by providing assistance in automatically summarized collocation relations drawn from multiple Japanese (sub-) corpora of different genres (Nishina 2011). An advantage of this is that comparison of the collocational behavior of words in different genres is provided. Further, the tool is convenient for searching synonyms and thereby exploring their similarities and differences through collocates. Recently, the tool has been provided in a set with the learner corpus Natane and Nutmeg, a system for automatic correction of learners' errors (Hodošček and Nishina 2012).

In 2010, the online corpus-query system NINJAL-LWP[6] was developed, initially for BCCWJ (Pardeshi and Akasegawa 2010), later incorporating the Tsukuba web corpus (TWC) (Imai et al. 2012). Similarly to Sketch Engine and Natsume, the tool provides lexical profiling, with a comprehensive picture of the collocational and

[3] Sketch Engine: http://www.sketchengine.co.uk/. The system is owned by Lexical Computing Ltd. One-month free trial is provided.

[4] Later on, another term appears for such types of tools, "lexical profiler". The name aims to grasp their advantages in search functionalities over standard concordance tools.

[5] Natsume: http://hinoki.ryu.titech.ac.jp/natsume/

[6] NINJAL-LWP: http://nlb.ninjal.ac.jp/

grammatical patterns of lexical entries. The system also visually compares the frequencies of search words in each of the subcorpora of BCCWJ, as well as the distribution of words within dialogues or within prose inside the book subcorpus.

With the compilation of the Balanced Corpus of Contemporary Written Japanese (BCCWJ), a new corpus-query tool Chunagon[7] was developed (Maekawa et al. 2013). This tool offers three major search functionalities for searching the morphologically annotated BCCWJ, namely the short-unit word, long-unit word and string search functionalities. By specifying a keyword and its surrounding context, the tool enables extraction of collocational relations as well. The extracted data can be downloaded in tabular format and then further analyzed within Excel or similar tools.

The corpus query tools introduced above have provided an essential contribution to the study of Japanese corpus linguistics and also have great potential for application in the fields of Japanese language education and second language acquisition.

3. Collocation data and possible applications in language education

This section focuses on adjectival and nominal collocational data and presents the results of the extraction of such collocations from Japanese language corpora. Based on the analysis of the obtained data, possible applications to Japanese language education are discussed, by comparing the results with the collocation data currently present in Japanese language textbooks.

4. "Collocation data of adjectives and nouns"

This collocation data resource was created for most frequent 500 adjectives in both, BCCWJ and JpTenTen. Once the most frequent adjectives were detected, for each of the adjectives their most frequent noun collocates were extracted using the Sketch Engine tool.[8] For highly frequent adjectives, up to 100 nominal collocates and for the rest of the adjectives up to 50 nominal collocates were extracted. In order to gain clean data, collocates below a frequency of 5 tokens in JpTenTen and a frequency of 2 tokens in BCCWJ were not included.

Both annotation data types, short-unit and long-unit words, were considered in this task, but since the JpTenTen with long-unit-word annotation is smaller in size and therefore fewer collocations can be obtained from it, long-unit-word annotated data was used only for the most frequent compound adjectives that were not obtainable from the short-unit-word annotated data. For example, *kyōmibukai* ‘interesting’ is annotated as

[7] Chunagon: https://chunagon.ninjal.ac.jp/

[8] Many thanks to the Sketch Engine team for technical support on this task.

one long-word-unit and two short-word-units: *kyōmi* ‘interest’ + *fukai* ‘deep’ and was therefore extracted from the corpus annotated with long-unit-words.[9]

One of the main drawbacks of obtaining collocation data from short-unit-word annotated data is that some units that are in general considered as a word are divided into two or more than two units reflecting their morphological derivation. For example, the word *kenkyūsha* is annotated as two short-word-units: *kenkyū* ‘research’ + *sha* ‘[a morpheme denoting a person]’, which can result in obtaining *kenkyū* when the actual use of the noun collocate was *kenkyūsha*. To overcome this issue, the three most frequent one-, two- and three-grams of word-units on the right and left of a collocate (right and left contexts of a collocate) were extracted and included in the final data, as shown in Table 1.

Table 1 exemplifies the obtained results of adjectives and their noun collocates along with their contexts. For the sake of space, only a few nouminal collocates of a few adjectives are displayed.

Finally, the obtained results based on large-scale resources and empirical methods clearly revealed the adjectives that have lexical constraints in their attributive role. For example *tebayai* “quick”, *tokorosemai* “crowded” are not used in their attributive adjectival form, preceding modified nouns, but rather in other forms, such as as an adverb preceding a verb (*tebayaku katazukeru* “to tidy up quickly”, *tokorosemashi to narande iru* “to line up crowdedly”).

5. Towards a dictionary of collocations for Japanese language learners

The resource “Collocation data of adjectives and nouns,” extracted from two large-scale corpora, can serve as a good basis for the creation of various resources that aim to describe Japanese i-adjectives and their behavior as modifiers of nouns. This section presents a number of major steps in analysis of the obtained collocation data, aimed at the creation of a dictionary of collocations for Japanese language learners. The analysis consisted of the following steps:

- Defining main entries for the dictionary

Srdanović (2013b) showed that the 25 most frequent i-adjectives represent more than 62% of the overall usage of i-adjectives in Japanese texts. Since our learner dictionary aims to cover very detailed information on adjective+noun collocates with thorough supplemental data, the initial version of the dictionary aimed only to cover these most frequent i-adjectives.

[9] For research on adjectives annotated as short- and long-word-units, refer to Srdanović (2013b).

Table 1: Example from the resource “Collocation data of adjective and nouns”

Ai freq.	Ai+N freq.	Left context 3+2+1	Left context 2+1	Left context 1	Ai	N	Right context 1	Right context 1+2	Right context 1+2+3	Word forms
3842021	33988	ま し た 。/の で 、/と 煙 は	の 一番/た 。/ま す 。	、/。/は	高い	所	に/から/で	に ある/が 好き/に いる	が 好き な/に ある の/から 落ち て	高い/高き/たかい
3842021	30911	と し て/世界 的 に/海外 で も	し て/で も/から も	で/も/に	高い	評価	を/が/と	を 得/を 受け/を 受ける	を 得 て/を 受け て/を 得 た	高い/たかい/高き
3842021	23027	は 非常 に/が 非常 に/、 そんな に	非常 に/そんな に/タダ より	は/に/、	高い	物	で/を/は	で は/の 、/で も	で は ない/に なっ て/が あり ます	高い/たかい/高き
3842021	13553	、 非常 に/は 、 より/は 非常 に	、 より/非常 に/は 、	より/に/、	高い	レベル	で/の/に	で の/に ある/で 実現	で 実現 し/で 安定 し/に ある 。	高い/たかい/高～い
3842021	13548	ま し た 。/の で 、/を 心臓 より	よ り も/た 。/は 、	、/より/も	高い	位置	に/で/から	に ある/で ボール/から の	で ボール を/に ある の/に あり 、	高い/たかい/高き
…	…	…	…	…	…	…	…	…	…	…
3808164	57103	。 これ は/だけ で も/いう の は	の は/これ は/本当 に	は/、/に	凄い	事	に/だ/です	に なっ/だ と/に なる	に なって/に なって る/だ と 思う	すごい/凄い/スゴイ
3808164	27036	ま し た 。/・ ・ ・/し て も	本当 に/た 。/に は	、/は/。	凄い	人	だ/が/です	だっ た/で し た/な ん	で し た 。/な ん だ/だっ た 。	すごい/凄い/スゴイ
3808164	17068	ま し た 。/！ 」 と/です が 、	と 、/て 、/が 、	、/は/。	凄い	勢い	で/です/の	で 食べ/で 、/で 走っ	で 食べ て/で 走っ て/で 売れ て	すごい/凄い/すっごい
3808164	16811	いう の は/は 本当 に/って の は	の は/と は/カ は	は/、/に	凄い	物	が/を/だ	が ある/が あり/です 。	が あり ます/が ある 。/で し た 。	すごい/凄い/スゴイ
3808164	6821	ま し た 。/て て 、/・ ・ ・	た 。/と か/て 、	、/。/も	凄い	格好	いい/よかっ/イイ	よかっ た/良かっ た/よく て	よかっ た 。/よかっ た です/よかっ た ！	すごい/すっごい/凄い
…	…	…	…	…	…	…	…	…	…	…
2931564	91894	いい こと と/は 何 も/良い こと も	の は/何 も/何 か	、/は/も	悪い	事	を/し/は	を し/で は/し た	で は ない/を し た/を し て	悪い/わるい/ワルイ
2931564	17591	良い ところ も/いい ところ も/良い ところ、	ところ も/ところ 、/ところ 、	、/も/。	悪い	所	は/が/を	が ある/は ない/も ある	で も ある/が あれ ば/が あっ た	悪い/わるい/ワルイ
2931564	16038	意味 で も/で も 、/ま し た 。	で も/も 、/は 、	も/、/。	悪い	意味	で/じゃ/の	で も/で は/で の	で はなく/で も 、/で は ない	悪い/わるい/惡い
2931564	14543	良い もの と/、 体 に/良い もの も	体 に/何 か/もの と	に/、/は	悪い	物	で/を/は	で は/で も/じゃ ない	で は ない/で は あり/で は なかっ	悪い/わるい/ワルイ
2931564	12627	良い 奴 、/いい 奴 、/は そんな に	本当 に/奴 、/そんな に	、/は/に	悪い	奴	ら/は/が	じゃ ない/で は/は い	で は ない/は い ない/、 変 な	悪い/わるい/ワルイ
…	…	…	…	…	…	…	…	…	…	…
2524026	24960	経つ の は/ま し た 。/たつ の は	の は/た 。/ま す 。	。/は/、	早い	物	で/です/勝ち	で 、/です 。/で も う	で す ね 。/で 、 もう/で 今年 も	早い/速い/はやい
2524026	17160	の で 、/、 比較 的/、 できる だけ	比較 的/は 、/、 かなり	、/は/かなり	早い	段階	で/から/に	で 、/で の/から 、	で 自分 の/で 出 て/で 気づい て	早い/速い/はやい
2524026	16108	いつ も より/の で 、/ま し た 。	に は/た 。/もう 少し	、/は/。	早い	時間	に/から/帯	帯 に/だっ た/帯 で	だっ た の/な の で/な の に	早い/速い/はやい
2524026	12239	、 できる だけ/の できる だけ/、 比較 的	できる だけ/比較 的/は 、	、/は/かなり	早い	時期	に/から/の	に 、/から 、/から の	に やって/です が 、/な の で	早い/はやい/速い
2524026	11785	の で 、/、 できる だけ/ま し た 。	で 、/は 、/できる だけ	、/は/。	早い	内	に/から/の	に 、/に 手/から 、	に 手 を/に やっ て/に 一 度	早い/はやい/速い
…	…	…	…	…	…	…	…	…	…	…

- Comparing collocations extracted from the two large-scale corpora

While we can use collocations from only one reliable large-scale corpus, there are some advantages in comparing the two large-scale corpora. The comparison reveals the idiosyncrasies of each corpus and helps us to exclude irrelevant collocates. On the other hand, results that are derived from both the corpora can be confirmed as significant. For example, the comparison of the noun collocates of the adjective *takai* 'high' in the two large-scale corpora of contemporary Japanese, BCCWJ and JpTenTen,[10] showed a very high level of similarity, which indicates the reliability of the corpora and the significance of the data (Srdanović 2013a: 149-51). A few differences among the first 50 collocates showed some more specific usages, such as *takai nobi/takai nobiritsu* 'high growth', which are often used in economics and specifically appear in the White book subcorpus of BCCWJ.

- Arranging the list of collocates

The list of noun collocates were arranged so that functional words such as *koto, mono, hazu, wake, tame* etc. are excluded. Then, the right and left contexts of collocates were checked to avoid the inclusion of incomplete collocates, and also in order to consider morphological elements that were excluded from the annotated short-unit word, for example, *(sei ga)takai kata* 'a tall person', *takai gijutsu(ryoku)* 'high skills/technology'.

- Grouping collocates based on their difficulty level

The twenty most frequent adjectives appear in the beginners' level of the old JLPT list and the Tsukuba vocabulary list, but they naturally appear in combination with their noun collocates at varying levels pf Japanese language study. Some of the collocates do not even appear in the JLPT list or the super advanced level of Tsukuba vocabulary list.

Grouping collocates into difficulty levels assists teachers and learners when considering collocates for inclusion at a specific level of language learning. Lower level collocates are not necessarily for inclusion at lower levels since in some cases a collocation requires a high level of proficiency even though its constituents are basic words. Learners' needs, motivation and other relevant factors also need to be considered when creating a foreign language learning syllabus. For example, while *takai* 'high/tall/expensive' and *kaimono* 'shopping' are both basic level words, their patterns of collocation are rather complex, abstract, and too specific to be included at the beginner level of a general Japanese language program. In contrast, higher level collocates would never be used in lower levels since their constituents are by definition difficult. This relationship between the level of difficulty of a collocate versus the difficulty of its constituents needs further clarification and calls for further collocation data analysis.

- Providing translations to a target language and furigana

[10] The comparison is done thus; the relative frequency of the first 100 nominal collocates for the adjective takai is calculated for each of the two corpora. The collocates are then sorted by the highest score of the sum of their relative frequencies in both corpora. Then, if a collocate appears in the list of the first 50 collocates in one corpus, but does not appear in the other corpus' list of the first 100 nominal collocates, it is specifically marked as different and further checked.

This study is limited to translating the obtained collocations into English, but can serve as a basis for preparing translations in other languages.[11] Besides English translations, the dictionary provides furigana in order to assist learners in reading unknown Japanese words.

- Detecting collocations that are difficult to predict or unpredictable for learners

Nation (2001) describes the so-called »unpredictable collocations« - collocations that are difficult for language learners to predict based on knowledge of their native or other foreign languages. Because a learner's prior knowledge is based on the constituents of collocations in their native language, if these constituents are different from those in semantically similar expressions in Japanese, the learner is likely to make a mistake and produce an utterance that is unnatural in Japanese. Therefore, it is important to grasp these types of collocations and pay special attention when introducing them to language learners by taking into account also how the collocates appear in their native language. The analysis of the collocates of *takai* (Srdanović 2013a) revealed a few different types of unpredictable collocates for learners of Japanese who are English native speakers. For example, *sei ga takai hito/kata* 'a high person' is unpredictable since the additional elements *sei ga* 'the back is' must be added to make the collocation complete in Japanese (lit. *a person with a high back*). Although the collocation *takai kabe* has the same constituent 'high' (*takai*) and 'wall' (*kabe*) as in English, the same expression has a second more abstract meaning 'high barrier' with different and unpredictable constituents. These unpredictable usages, different from expressions in learners' mother tongues, need to be explicitly explained. Finally, it is important to note that unpredictability is relative to the language background of learners and needs to be considered separately for each language.

- Discovering complex patterns, typical usages, genre specifics etc.

Further, the corpus-based analysis reveals complex patterns, typical usages, and usages specific to a particular sub-corpus/genre, which are then described in the learners' dictionary. This type of information is specifically searched for in the case of unpredictable collocations. The following is an example of a detailed description on differences in usage between *takai koe* 'high-pitched voice'and *ōkī koe* 'loud voice'.

高い声 「高い声」と「大きい声」は、意味・用法が違う。

「高い声」＋で＋「歌う、鳴く、叫ぶ」（with a high-pitched voice）

「大きい声」＋で＋「言う、話す、歌う、叫ぶ、挨拶する」

（with a loud voice, loudly）

- カラオケで高い声 がでなくなって辞めた 。｜若いメイド達が悲鳴の様な高い 声を上げた。｜

[11] This could possibly be a basis for theoretical investigations into the cross-linguistic behavior of collocations.

以前 出ていた高い声が出なくなったらどう思いますか？

- 大きい声で言ってください。（注：このよく使われる表現は「高い」で使えない。）｜

大きい声では言えないような話も多い。

- Providing lexical maps of noun collocates

Finally, the dictionary provides images of lexical maps of noun collocates, which depict how collocates can be grouped together as concepts observed from a cognitive linguistics perspective. The analysis of *takai* and usage of its noun collocates revealed three major groups of collocates: collocates that represent relative position (e.g. *takai yama* 'high mountain', *takai ki* 'high tree', *takai tokoro* 'high place'), relative quantity (*takai wariai* 'high percentage', *takai nedan* 'high price') and relative quality (*takai shinraisei* 'high reliability', *kanshin* 'high/great interest').

For more information on how entries are structured in the dictionary, refer to Srdanović (2013a). This study introduces the adjective *takai* and its noun collocates as an example from the ongoing dictionary project.

6. Evaluation of Japanese language textbooks for their collocation data

This section summarizes the results of the analysis of the adjective *takai* and the use of its nominal collocates in the beginner Japanese language textbooks *Minna no Nihongo honbun* 1 and 2. The results are compared to the collocation data obtained from the corpora and processed for the target learners' collocation dictionary described in the previous section. The observations are as follows:

- The adjective *takai* is introduced in the beginner textbooks in various grammatical forms and patterns including the attributive form (*rentai*). This adjective is presented in its predicative form (*shūshi*) more than twice as often than in its attributive form. The behavior of the adjective in the corpus indicates, however, that these two forms of *takai* have approximately the same distribution.[12]
- *Takai* 'high, tall' in its attributive form is introduced with nouns such as *yama* 'mountain', *biru* 'building' and in the pattern *sei ga takai hito* 'a tall person', which only covers only the semantic domain of relative position. The use of corpora revealed two other major domains (relative quantity and relative quality) covered by the collocates of *takai* in its attributive form.
- Among the collocates not covered in the textbooks are: *tokoro* 'place', *tatemono* 'building', *ki* 'tree', *kaimono* 'shopping', (*o*)*kane* 'money' which belong to JLPT

[12] I presented the analysis of the distribution of forms of the adjective takai from JpTenTen at the 27th Paris Meeting on East Asian Linguistics, at CRLAO / INALCO. They are summarized as: takai+N (18%), takai+suffix (18%), N+takai (10%), Nga/no+takai+N (16%), takaku+V (9%), Nga+takaku(te)[cont] (8%), Nga+takai[concl] (18%), Nwa+takai[concl] (3%).

4th level and correspond to a low beginner level. Also, *oto* 'sound', *kabe* 'wall/barrier', *basho* 'location', *wariai* 'percentage', *gijutsu(ryoku)* 'technology/skill', *kyōiku* 'education' belong to JLPT 3rd level and correspond to an upper beginner level.[13]

- Some of the above mentioned collocates are less predictable by Japanese language learners and therefore require special treatment. For example, *sei ga takai hito* 'a tall person' is not correct if used only with the constituents *takai hito* 'lit. tall + person]'. Learners make such mistakes, as noted by Srdanović and Sakoda (2013) and therefore Japanese language learning materials need to make leaners aware of this.

To summarize, there is a need for more systematic treatment of collocations in textbooks, in relation to forms, relevant semantic domains, levels of difficulty and predictability. Also, a well-balanced large-scale spoken corpus, once created for Japanese, would be a valuable resource for further analysis.

7. Conclusion

This paper introduced some major tools, resources and methods that can be used in analysis of collocations. "Collocation data of adjective and nouns", which was obtained from two large-scale Japanese language corpora has been described. The resource is valuable as an exhaustive collection of the collocational data of i-adjectives and nouns and can be used as a basis for the creation of various empirically based materials that would aim to describe Japanese i-adjectives when they act as nominal modifiers. Also discussed was a more in-depth analysis of "Collocation data of adjective and nouns" aimed at the creation of a dictionary of collocations for Japanese language learners. Finally, the comparison of the data with the collocation information in textbooks revealed how the material could be improved based on the data obtained from corpora.

The Japanese language corpora and tools, collocation resources, research methods and results that were introduced in this paper will hopefully contribute to future creation of corpus-based Japanese learners' dictionaries, textbook materials and syllabi.

Literature

Fukada, A. (2007) Chakoshi: a Japanese text search and collocation extraction application [in Japanese]. In: *Japanese linguistics 22*, 161-172.

Himeno, M. (ed.) (2012) Kenkyusha Japanese Collocation Dictionary (Kenkyūsha Nihongo Korokēshon Jiten). Tokyo: Kenkyusha.

[13] Some of the items, although belong to beginner level, are more demanding to master as collocations and hence not needed at this level, for example takai kaimono 'expensive purchase', takai okane 'lots of money'.

Hodošček, B. and Nishina, K. (2012) Japanese Learning Support Systems: Hinoki Project Report. *Acta Linguistica Asiatica 2(3).* Ljubljana: Ljubljana University Press, 95-124.

Imai, S., Akasegawa, S., Pardeshi, P. (2013) Development of NLT: the Search Tool for Tsukuba Web Corpus, *Proceeding of the 3rd Japanese corpus linguistics workshop*, Department of Corpus Studies/Center for Corpus Development, NINJAL, 199-206.

Kawahara, D. and Kurohashi, S. (2006) Case Frame Compilation from the Web using High-Performance Computing. In*: Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC2006)*.

Kilgarriff, A., Rychly, P., Smrz, P. and Tugwell, D. (2004) The Sketch Engine. In: G. Williams and S. Vessier (eds) *Proceedings of Euralex*, 105–16. Bretagne, France: Université de Bretagne-Sud.

Maekawa, K., Yamazaki, M., Ogiso, T., Maruyama, T., Ogura, H., Kashino, W., Koiso, H., Yamaguchi, M., Tanaka, M. and Den, Y. (2013) Balanced corpus of contemporary written Japanese. *Language Resources and Evaluation*. Netherlands: Springer.

McEnery, T. and Hardie, A. (2012) *Corpus Linguistics: Method, Theory and Practice*. Cambridge Textbooks in Linguistics. Cambridge University Press.

Nation**,** P. (2001) *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Nishina, K. (2011). Development and evaluation of the writting support system Natsume using, the balanced corpus (in Japanese). *Tokutei ryōiki kenkyū »Nihongo kōpasu« Heisei 22 nendo kōkai wākushoppu yokōshū.* Tokyo: Monbukagakushō kagakukenkyūhi tokuteiryōiki kenkyū 'Nihongo kōpasu' sōkatsu ban. 215-224.

Oso M and Takizawa, N. (2003) Corpus-based study on Japanese education: About collocations and their misuse (in Japanese). *Japanese Language* 22(5). Meiji shoin, 234-244.

Pardeshi, P. and Akasegawa, S. (2010). BCCWJ wo katsuyō shita kihon dōshi handobukku sakusei: kōpasu braujingu shisutemu NINJAL-LWP no tokuchō to kinō. *Tokutei ryōiki kenkyū Nihongo kōpasu Gendai Nihongo kakikotoba kinkō kōpasu kansei kinen yokōshū*. 205-216.

Srdanović, I. (2013a) Description of Adjective and Noun Collocations Based on Large-Scale Corpora: Towards Dictionary for Japanese Language Learners (in Japanese). *Kokuritsu kokugo kenkyūjo ronshū (NINJAL Research Papers)* 6.

Srdanović, I. (2013b) Japanese i-adjectives as short and long-word units: implications for language learning. In: *Proceedings of the Conference of the Pacific Association for Computational Linguistics (PACLING),* 8 pp (CD-rom).

Srdanović, I. and Sakoda, K. (2013) Analysis of learner's production of adjectives using the Japanese language learner's corpus C-JAS: the case of *takai*. *Acta Linguistica Asiatica*, 3(2), 9-24.

Srdanović, I., Erjavec T. and Kilgarriff, A. (2008) A web corpus and word-sketches for Japanese. *Shizen gengo shori (Journal of Natural Language Processing)* 15(2), 137-159.

Srdanović, I., Suchomel, V., Ogiso, T, Kilgarriff, A. (2013) Japanese Language Lexical and Grammatical Profiling Using the Web Corpus JpTenTen (in Japanese).In: *Proceeding of the 3rd Japanese corpus linguistics workshop*. Tokyo: NINJAL, Department of Corpus Studies/Center for Corpus Development, 229-238

CONSTRUCTION OF A LEARNER CORPUS FOR JAPANESE LANGUAGE LEARNERS: NATANE AND NUTMEG

Kikuko NISHINA
Tokyo Institute of Technology
nishina.k.aa@m.titech.ac.jp

Yagi YUTAKA
Picolab Co., Ltd.
yagi@picolab.jp

Bor HODOŠČEK
Osaka University
bor@lang.osaka-u.ac.jp

Takeshi ABEKAWA
National Institute of Informatics
abekawa@nii.ac.jp

Abstract

Japanese language learners aim to acquire reading, listening, writing and speaking skills. We at the Hinoki project (https://hinoki-project.org/) have recently been working on the Natsume collocation search system (https://hinoki-project.org/natsume/), the Natane learner corpus to support Natsume (https://hinoki-project.org/natane/) and the Nutmeg writing support system (http://hinoki-project.org/nutmeg/). In order to test the effectiveness of Nutmeg, we conducted an online experiment with 36 participants who used the system's register misuse identification feature to correct four writing assignments. Results show that Nutmeg can be an effective tool in correcting common register-related errors, especially those involving auxiliary verbs. However, the accuracy of verb and adverb identification was too low, suggesting the need for improvements in the variety of corpora used for identifying register misuse.

Keywords: writing support system; learner corpus; academic writing; register; errors; performance evaluation experiment

Povzetek

Cilj vsakogar, ki se uči tuj jezik, je, da usvoji branje, slušno razumevanje, pisanje in govorne sposobnosti ciljnega jezika. S projektom Hinoki (https://hinoki-project.org/) si prizadevamo narediti iskalnik kolokacij Natsume (https://hinoki-project.org/natsume/), učni korpus Natane, ki bo podpiral Natsume (https://hinoki-project.org/natane/) in podporni sistem Nutmeg za pisanje (http://hinoki-project.org/nutmeg/). S spletnim eksperimentom, ki je vključeval 39 sodelujočih, smo ocenili učinkovitost sistema Nutmeg. Vsak sodelujoči je s pomočjo uporabe identifikacijskih lastnosti za napačno uporabo jezikovnega registra, ki jih ponuja sistem Nutmeg, popravil štiri pisne naloge. Rezultati kažejo, da je sistem Nutmeg učinkovito orodje za popravljanje splošnih napak, ki so povezane z registrom jezika, še posebej v primerih pomožnih glagolov. Hkrati smo ugotovili, da je prišlo do nepravilnosti pri prepoznavanju glagolov in prislovov, zaradi česar bo

Acta Linguistica Asiatica, 4(2), 2014.
ISSN: 2232-3317, http://revije.ff.uni-lj.si/ala/
DOI: 10.4312/ala.4.2.37-51

potrebno povečati raznolikost korpusov, na katerih prepoznavamo napačno uporabo jezikovnega registra.

Ključne besede: podporni sistem pisanju; učni korpus; akademsko pisanje; register; napake; eksepriment ocenjevanja uspešnosti

1. Introduction – the aims of developing learner support system

Japanese language learners aim to acquire reading, listening, writing and speaking skills. We at the Hinoki project (https://hinoki-project.org/) have been working since the mid 1990's to develop Asunaro, a Japanese language reading comprehension support system, and by 2010 had created the Natsume collocation search system (https://hinoki-project.org/natsume/), the Natane learner corpus to support Natsume (https://hinoki-project.org/natane/) and the Nutmeg writing support system (http://hinoki-project.org/nutmeg/) (Hodošček & Nishina, 2012). Project members included Takeshi Abekawa, Yutaka Yagi and Kikuko Nishina, and were later joined by Bor Hodošček in 2008 (Nishina et al., 2012). In this paper, we will first present an overview of Natsume, followed by Natane and Nutmeg's research aims and current state of progress, and finally evaluate the system's effectiveness by conducting a performance evaluation experiment.

Natsume allows users to search for other words that co-occur with the word they want to use and provides hints for constructing sentences (Hodošček et al., 2011). However, it cannot directly contribute to the formation of correct sentences by correcting mistakes in the user's input. Another shortcoming is that although advance level learners can easily operate Natsume's user interface, learners belonging to the intermediate levels or below face difficulties in fully utilizing the system. It is not enough to simply provide the correct usage of a vocabulary item to such learners. They also need a system that points out incorrect word usages within sentences along with suggestions on how to correct them. To realize this we need to create a Japanese writing support system like Nutmeg which statistically determines the appropriateness of a term or a usage through Natsume (which contains authentic Japanese corpora) and Natane (a corpus of learner's compositions). In this paper, we present the results of our experiment that utilized learners' academic reports and papers to evaluate the effectiveness of Nutmeg.

2. Natsume – a Japanese collocation search tool

Natsume makes use of the Balanced Corpus of Contemporary Written Japanese (BCCWJ) by the National Institute of Japanese Language and Linguistics (NINJAL), the Japanese version of Wikipedia and a corpus comprising of scientific papers compiled independently by the Hinoki Project. The system generates a list of collocations of the word specified by the user, which can be sorted through a variety of statistics. The user may simply type in the word (a noun, verb or adjective) they wish to find collocations

of, click 'search' and the system generates a list. Clicking on a collocation set will additionally display the collocation's frequency distribution among several genres and allow the user to view example sentences. As can be inferred from this, Natsume is a useful tool for advanced learners, but intermediate learners may find the process of selecting an appropriate word, searching through example sentences to infer its meaning and then judging whether the selected word can be successfully incorporated into the sentence, too challenging.

Table 1: Learners and the number of compositions

Mother-tongue	Number of learners				Number of compositions			
	male	female	no respons	total	male	female	no response	total
Chinese	50	43	22	115	62	64	26	152
Marathi	6	23	7	36	6	23	7	36
Vietnamese	6		7	13	18		9	27
Korean	6	1	4	11	24	3	7	34
Spanish	2			2	2			2
Malay	1			1	8			8
Slovene	1			1	7			7
Hungarian	1			1	1			1
Thai			1	1			1	1
No response	1		10	11	5		12	17
Total	**74**	**67**	**51**	**192**	**133**	**90**	**62**	**285**

3. Natsume – a corpus of learners' compositions

As was pointed out above, a function that corrects sentences constructed by learners has to be added in order to create a writing support system geared towards the needs of intermediate learners. This entails compiling a learner corpus of Japanese compositions, analyzing the frequently occurring errors found in the data and identifying their causes in order to create a system that automatically points out and corrects learner mistakes. In order to achieve this, we have compiled Natane by independently collecting learners' compositions since 2011 and asking Japanese language teachers to tag the errors. The error tags comprise of "error item", "error content" and "cause of error". These are further divided into three hierarchies and consist of approximately 70 varieties (Cao et al., 2012).

3.1 Collecting learner compositions

As of August 2014, the corpus consists of 285 compositions (with a total character count of 205,520) written by 192 learners belonging to graduate and undergraduate programs of different universities or foreign language schools. The compositions include

approximately 6,500 errors that have been tagged using around 9,000 errors tags. It also includes metadata like sex, nationality, mother tongue, period of studying Japanese, and Japanese language proficiency wherever this was possible. Table 1 gives the breakdown by mother tongue and sex. Approximately 60% of the learners are Chinese. Marathi[1], Vietnamese and Korean learners make up around 31% and the remaining 5 groups make up 9% of the total number of learners. The ratio of countries that use Chinese characters in their writing system vis-à-vis those that do not is six to four if we do not include Korea in the former group. The data is therefore slightly biased. The average number of characters per composition is around 720 words.

3.2 Designing the error tags

The next step was to decide how to apply the error tags in the compositions. Hideo Teramura's analysis on learner errors (1990) is a well-known reference work of over 430 pages consisting of 7415 sentences written by learners. The data is organized into four columns – 1) sequential number, 2) learner's nationality, composition type, 3) the sentence containing the error(s) and 4) error type. Error types are divided into five categories – phonology (pronunciation), writing system (script), lexicology, morphology (conjunction), and syntax & semantics. Examples of erroneous sentences are arranged in order of error types starting with phonology. Yasuko Ichikawa's study (1997) builds on Teramura's research. Besides these, corpora like Shanghai Jiao Tong University's Learner's corpus of written Japanese (http://tesol.sjtu.edu.cn/corpus/index.php/Public/login) and Tokyo University of Foreign Studies' Learner's Language Corpus of Japanese (http://cblle.tufs.ac.jp/llc/ja/index.php) are also available on the web. These corpora were released around the same time as Natane.

Drawing on insights from previous error tagging efforts, we analyze each error from three different viewpoints, namely the "error item", the "error content" and the "cause of error". "Error target" lists the places where errors occur. "Error content" gives an analytical account of the error from the perspective of discourse, construction, vocabulary, phonology and script. "Cause of error" indicates whether the error is due to mother tongue-influence, confusion in meaning, pronunciation or character shape etc. These reference frames are further organized into three hierarchies, and the total number error types is seventy (Cao et al., 2012).

Table 2 presents the frequency of error tags found in "cause of error" arranged by learner's mother tongue. The alphabets in the top row represent the learner's mother tongue and the rows below display error tag frequency. The bottom row gives the total frequency found in Natane. Confusion with similar expressions is sub-categorized into similar meaning, similar character and similar sound.

[1] Marathi belongs to the Indo-European group of languages. It is an official language of India, spoken mainly in the western state of Maharashtra.

Table 2: Causes of errors

Items	Zh	Mr	Vi	Ko	Es	Ms	Sl	Hu	Th	NR	Total
similar meaning	38	141	11	32		2	7		2	10	243
similar character	2	47	1	4		2	1				57
similar sound	7	110	1	10		3	2		1		134
mother-tongue influence	45	6	1	5				1			58
register	384	12	8	46	9		2	4		18	483
miss-match in styles	411	21	10	10	9			3		14	478
other	12	3	1	3						2	21
Total	**899**	**340**	**33**	**110**	**18**	**7**	**12**	**8**	**3**	**44**	**1474**

zh: Chinese, mr: Marathi, vi: Vietnamese, ko: Korean, es: Spanish, ms: Malay, sl: Slovene, hu: Hungarian, th: Thai, NR: no response

3.2.1 Similarity in character and sound

In general, the similarity in sound and that in character are inter-related, making it difficult to judge the cause. Examples include such errors as *ishhou ni* (一生に, for a lifetime) instead of *issho ni* (一緒に, together) and *yuumei* (有名, famous) instead of *yume* (夢, dream). Learners from countries that use Chinese characters and advanced learners frequently use Chinese characters making it difficult to detect whether they have correctly acquired Japanese phonemes. However, learners who are from countries that do not use Chinese characters tend to rely on the Kana syllabary, making errors due to similarity in pronunciation and character more pronounced. This contrasts sharply with errors concerning similarity in meaning, which are observed among advance level learners or learners from countries that use Chinese characters.

3.2.2 Mother tongue influence

The total number of errors attributed to mother tongue influence in the entire corpus was only 58. Approximately 80% of these were errors in selecting Chinese characters by Chinese learners. For example, the Chinese word 階段 corresponds to 段階 (phase) in Japanese and an error where 階段 (meaning ‘staircase’ in Japanese) was used instead of 段階 (phase) was observed.

3.2.3 Register

The term “register” is defined along the same lines as that in the systemic functional grammar (SFG) framework proposed by Michael Halliday and associates. In SFG, differences in linguistic expressions are described as a set of linguistic options available in a particular language setting which have certain social restrictions placed

upon them, a concept referred to as "register" (Halliday & Hasan, 1976; Biber & Conrad, 2009). This means that different vocabulary and grammar items may be used depending on the relationship between the writer and the reader and the context within which the exchange takes place. In the case of learner compositions, using spoken expressions in a class report is inappropriate to the context and therefore translates as an error concerning register. At present there are around 483 error tags concerning register, over half of which are errors concerning difference among written and spoken expressions.

Figure 1: Natane's search interface

A tendency to use expressions learnt in conversation class for beginners can be observed in learners' academic writing. Examples include the use of sentence-final particles such as "ne", which is typically used to ask for confirmation or consent during conversations, particles such as "toka" and "-shi", conjunctions such as "demo" and "dake", sentence-final subsidiary verbs such as "chau" and other verbs such as "yaru". Furthermore, learners may also use subjective expressions like "sei" which is a conjunction that conveys causal relationship. These have been classified as register related errors because using such an affective style is inappropriate in a formal report. Beginner level learners were excluded when considering errors concerning register. The reason for this was that the scope of their vocabulary and expressions was mainly limited to spoken Japanese expressions and lacked other variations. Register becomes more of a problem once the learner has reached the intermediate or advance levels and is required to use academic language to write reports. Besides the ability to distinguish between written and spoken language, the ability to maintain one style throughout the composition is also necessary.

At present, there are no educational materials that systematically teach that this distinction is a difference between two registers. This leads us to surmise that we lack sufficient teaching material and courseware geared towards teaching academic expressions for advanced learners and the development of new material is necessary. For

this reason, the remainder of this paper will focus on the acquisition of registers within the scope of conducting the performance evaluation experiment in Section 5.

3.3 Natane – a tool for searching errors

The learner corpus Natane thus compiled is equipped with search functions. By inputting conditions such as parts of speech, specific vocabulary items or mother tongues, one can look-up examples of errors situated within the complete text (https://hinoki-project.org/natane). Figure 1 shows a part of the search interface of Natane. Using this tool, Japanese language teachers can find out what the common errors are and may even use it as reference material while planning classes or preparing exercises.

4. Nutmeg – writing support system

As will become clear from the survey results discussed below, learners want an error correction system in a writing support system. An example of related research targeting Japanese native speakers was JSS, an automatic grading system for short theses (http://coca.rd.dnc.ac.jp/jess/; Ishioka, 2008). The system promptly grades the text's rhetoric, ratio of Chinese characters, number of embedded sentences, diversity in vocabulary, logical structure, sentence construction, content, and length on a 10-point scale. However, it does not indicate problematic sections within the text, nor does it guide the user on how to correct them. In the case of a writing support system for non-native speakers, it is important to specify the error location as well as the reasoning that underlies their correction. For this reason we conceptualized Nutmeg: a system which not only gives a feedback on necessary corrections, but also gives the reasons for suggesting those corrections.

Figure 2 shows Nutmeg's writing correction screen. Learners can write or paste in their writings and press the "添削" (correct) button, upon which the system marks the errors. Clicking on the marked word or expression makes a pop-up appear displaying reasons for the error like "register error" and learners may follow the suggestion to make corrections. The technique proposed by Hodošček & Nishina (2011) is used to give feedback on errors. In this technique, register related errors are determined by using data from the BCCWJ (a corpus of written Japanese consisting of a variety of registers like books, magazines, newspapers etc.) and a corpus of scientific papers. We assume the register learners intended for their report and set-up a quasi-correct data, which is close to the target data, and a quasi-incorrect data, which is distant from the target data from the various BCCWJ registers and the corpus of scientific papers. After that, we conduct a chi-square test based on the distribution of co-occurring expressions that patterns as combinations of "noun＋particle＋verb", "noun＋particle＋adjective" and "adjective ＋noun". In case the frequency in the quasi-incorrect data is significant, the co-occurring

expression is labeled as inappropriate under the target register. The data was further expanded in Yagi et al. (2014), where register related errors and independent morphemes could be detection by an arbitrary 3-, 2-, or 1-gram.

5. Performance evaluation experiment for Nutmeg

5.1 Method

In order to test the effectiveness of Nutmeg, we conducted the following tests, surveys and experiment from January 2014 until April 2014 (Yagi et al., 2014). All tests and surveys were performed online. The overall focus of the experiment was to evaluate the effectiveness of the error detection function for errors related to register.

1) J-CAT (Japanese Computerized Adaptive Test) (Imai & Kuroda, 2012)
2) Participants background survey
3) Report writing assignments on four topics using Nutmeg
4) System survey

Before conducting the main experiment we checked the current Japanese language proficiency level and linguistic background of all participants and asked them to take J-CAT. J-CAT results are divided into seven levels, as shown in Table 3. The levels from 'lower-advanced' to 'beginner' correspond to levels 1 to 4 of the old Japanese Language Proficiency Test (JLPT). J-CAT consists of a listening, vocabulary, grammar, and reading-comprehension test. Next we conducted the survey on participants' background. This was followed by the main test – asking the participants to write a report of 400 characters or more on the following four topics:

Assignment 1: "Things I do not understand about Japanese people"
Assignment 2: "Pros and cons of nuclear electric power generation"
Assignment 3: "Reasons for the popularity of Japanese anime"
Assignment 4: "Merits and demerits of the Internet"

Assignments 1 and 3 were selected as familiar issues that even participants with somewhat lower proficiency levels could write about, whereas assignments 2 and 4 required that the argument be developed in a logical manner. The participants were first presented with the topic and prompt, and once they completed the report they were asked to make changes based on the corrections suggested by the system's register error detector. Participants could make changes based on the feedback only once in order to simplify the process of comparing the reports before and after corrections, as well as to prevent participants from iteratively correcting without thought. Also, a minimum of a three day gap was placed between each assignment in order to note the changes in learning effect. After the main test was completed, participants were asked to fill in a

questionnaire comprising of six questions, a section asking the names of dictionaries they used and another section asking them to freely express their opinions about the error detection system.

Table 3: Proficiency level of participants

proficiency	J-CAT	JLPT[1]	distribution(name)
Near-native speaker level	Above 350		1
Advance level	300－350		7
Lower-advance level	250－300	Level 1	17
Upper-intermediate level	200－250	Level 2	7
Intermediate level	150－200		4
Lower-intermediate level	100－150	Level 3	0
Beginner level	Below 100	Level 4	0

Figure 2: Interface for the Nutmeg performance evaluation experiment

Table 4: Average composition length

	Assignment 1	Assignment 2	Assignment 3	Assignment 4
Avg. length	545.94 characters	508.53 characters	511.08 characters	523.90 characters

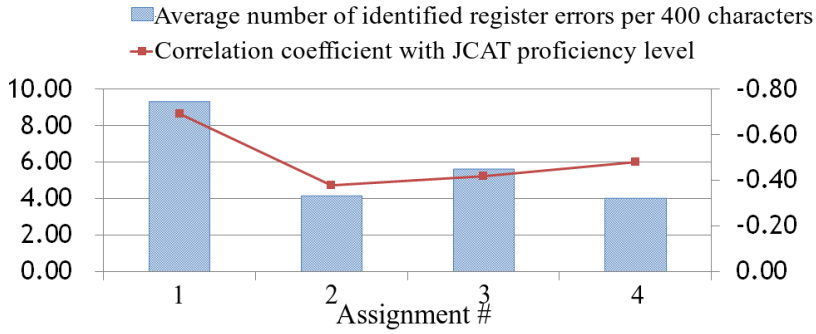

Figure 3: Average number of register related errors identified per 400 characters

Figure 2 shows the interface displaying the correction result. The bottom left of the screen displays the results of the corrections. Potential errors are highlighted and underlined. Participants can click on the indicated erroneous word or expression to read the suggested change and make corrections in the editing box at the bottom right of the screen after checking the correct usage in a dictionary or through Natsume, if required. Words and expressions identified by the system and the content of the corrections made by the participants, a writing log tagged with elapsed time taken at 10 second intervals, sections that were clicked to confirm the correction content etc. are all logged in the system database backend and can be used to analyze participant behavior.

The methodology includes a description of participants, stimuli and procedure used to study Hindko plosives acoustically.

5.2 Participants

The experiment was conducted with the cooperation of 36 participants from three universities in Japan and two foreign universities. Table 3 gives the break-up of the participants' proficiency level. The average age was 25.3 years. Fifteen participants were under-graduate students and twenty-one were graduate students. The proportion of male and female participants was 11 male participants and 25 female participants. There were 26 Chinese, 4 Korean, 4 Slovenian and 2 Croatian participants. The J-CAT results reveal that in terms of the can-do statements regarding writing skills, 50-75% of advance learners could write a report on a field of their interest and 25-50% can present their arguments or opinions in a logical manner. Among the intermediate learners, 50-75% could write a report, and 25-50% could present their arguments or opinions in a logical manner. As can be seen from this survey, all the participants in this experiment were acceptable candidates for carrying out the main assignment of "writing a report".

6. Results and observations

As the participants were asked to write their reports at any time convenient to them, it took around three months after the instructions were first published on the experiment website to collect all the reports.

6.1 Correlation between the number of errors identified by the system and participant proficiency

Table 4 gives the average length of reports for each assignment. Assignment 1 was the longest and assignment 2 the shortest. Figure 3 is a bar graph representation of register related errors identified by the system normalized to occurrences per 400 characters. The number of identified errors per assignment was not correlated with text length, nor was text length correlated with participant proficiency levels. The line graph shows that there is an inverse correlation between the number of errors identified by the system and the J-CAT results. The fact that the value is comparatively much greater in the case of assignment 1 than the subsequent assignments 2, 3 and 4 indicates that this may be due to the existence of a learning effect that took place after the system identified the register related errors in the first assignment, which will be explored in the following analysis.

6.2 Participants' responses to the corrections suggested by the system

The register related corrections suggested by the system are organized by their part of speech in Table 5 along with the assessments by Japanese language teachers regarding their validity. The highest number of corrections (by token frequency) was suggested for auxiliaries followed by adverbs and verbs. On the other hand, the category where the largest number of invalid corrections was made was verbs, followed by adverbs and particles.

The token frequency of valid corrections made in the category of verbs was 86 as opposed to 97 invalid corrections. The selection of an appropriate word-group for this category, in particular, was a trial and error process to begin with, and irrespective of the fact that the threshold was adjusted, it was difficult to reach a stable result. The type frequency of the 97 invalid verbs is 21 verbs that include *ieru* (to be able to say), *oshieru* (to teach), *sagasu* (to search), *nayamu* (to worry), *kaeru* (to return), *gozaru* (to be) (in the order of their frequency). The verbs *iu, oshieru* and *kaeru* are level four vocabulary items in the old JLPT and the majority of the verbs are beginner level vocabulary items. On the other hand, the statistical tests conducted on the corpora confirm that their use within the quasi-correct data is significantly low when compared to the quasi-incorrect data. The corpus for scientific writing, which contains expressions that appear in academic reports, has its limits and is not exhaustive. One issue is that Japanese language learners tend to use beginner level vocabulary as they have not yet acquired typical

expressions used in academic reports. This makes it necessary to allow alternative wordings using more basic words as part of academic writing, even though these words and expressions may not be typical in research paper writing.

Table 5: Validity of the corrections

Part of speech	valid	invalid	total	Accuracy
Auxiliary verb	342	0	342	100.00%
Adv erb	158	29	187	84.49%
Verb	86	97	183	46.99%
na adjective	88	11	99	88.89%
i adjective	83	12	95	87.37%
Particle	55	21	76	72.37%
Noun	42	4	46	91.30%
Subsidiary verb	10	0	10	100.00%
Adnominal adj	10	0	10	100.00%
Interjections	3	5	8	37.50%
Total	**877**	**179**	**1056**	**83.05%**

Table 6: Participants' responses to the corrections

Part of speech	no change	changed	deleted	deleted%
Auxiliary verb	258	25	59	14.91%
Adverb	107	51	29	15.51%
Verb	138	39	6	3.28%
na adjective	69	23	7	7.07%
i adjective	62	31	2	2.11%
Particle	51	21	4	5.26%
Noun	33	11	2	4.35%
Subsidiary verb	4	1	5	50.00%
Adnominal adj	7	3	0	0.00%
Interjections	6	0	2	25.00%
Total	**735**	**213**	**108**	**10.23%**

The second highest number of invalid corrections was made for the adverb category. The type frequency for adverbs was 24 words, out of which 6 words, that is, *sara ni* (further), *kanarazu* (always), *mushiro* (rather), *mattaku* (entirely), *touzen* (completely), *juubun* (sufficient) are not included in the register for academic papers and reports. However, it is necessary to acknowledge these words as part of the register for academic writing since they are found in this register.

Table 6 presents the actual numbers drawn from the user activity log data where items that were not changed by the participants are labeled "no change", items that were changed are labeled as "changed" and items that were deleted are labeled "deleted".

As seen from table 5, the system sighted auxiliaries as the part of speech that had the most inappropriate usages, out of which 335 were concerning use of the *desu–masu*

form, occupying 97% of the total. Two thirds of these were left unchanged and in the case of the remaining one third, either the "verb+*masu*" pattern was changed to the basic verb form and the "*desu"* pattern to the "*~de aru*" form or they were deleted altogether. Adverbs, the next most frequent after auxiliaries, numbered a total of 187 words out of which 107 were left unchanged, 51 were changed and 29 deleted. Adverbs like *sara ni*, *mushiro*, *mattaku* and *kanarazu*, which had been deemed inappropriate by Japanese language teachers by the same method as mentioned previously, had a token frequency of 27 words and were all left unchanged by participants.

Furthermore, other adverbs that were left unchanged such as *hakkiri* (clearly), *taihen* (very), *yappari* (after all), *chotto* (slightly), *chanto* (properly), *takusan* (a lot), *sou* (that), *zutto* (throughout), *iroiro* (various), *kichinto* (orderly) etc. had a token frequency of 80. The only instance of a "change" that was an improvement was replacing *motomoto* (originally) with *honrai* (originally). Other changes that cannot be considered as improvements included changes such as replacing *iroiro* (various) with *takusan* (a lot) and *zenzen* (very/entirely) to *sugoku* (very/extremely). At present, there is no means to prevent such inappropriate changes from being suggested by the system. However, adverbs used in academic writing exclude emotional and sentimental expressions, and are limited in number. It is therefore possible to form a list of adverbs used in academic writing and this is one improvement left for future work.

Many of the "deleted" words were spoken expressions, but there were also words like *mushiro*, *kanarazu* and *juubun* that the Japanese language teachers deemed as inappropriate changes suggested by the system. We may think of these as words deleted by participants due to their inability to come to a definite conclusion. The results show that not only do the participants have little knowledge about adverbs used within the register of academic writing but also that the "defects" in the data are a source of confusion. The experiment clearly shows that the corpus of reports and academic papers has several drawbacks that require improvements.

6.3 Survey results

After the main experiment, we conducted a questionnaire asking the participants to rate the system. Though most participants gave it a high rating, they also pointed out some problems. Below we present the comments written by participants in the 'comment freely' section of the questionnaire. Among the positive evaluation we find comments such as, 1) I realized my mistakes thanks to the suggested corrections, 2) I learnt that the expressions I had used are not employed in reports, and 3) I found out my mistakes concerning particles used before and after sentences and patterns. The following areas for improvement were also suggested – 1) "it was unfortunate that there wasn't any information given as to why the indicated expressions are inappropriate. I would have like a more detailed explanation on how to correct it.", 2) (regarding Natsume) "since information is organized according to parts of speech, it becomes difficult to search for

expressions if one does not know their part of speech", 3) (Natsume) "...has very few synonyms and the example sentences were difficult to understand".

Regarding the first comment in areas that need improvement, the reason for withholding the correct usage and providing reference information instead was to allow participants to think for themselves. However, it may be necessary to provide scaffolding to intermediate level learners in order for them to come to the level where they can correct their own mistakes. The other two comments are related to the user experience in Natsume, and require further investigation.

7. Conclusion

The experiments show that Nutmeg is an effective tool. On the other hand, we also discovered problems with regards to verbs and adverbs that are omitted from the register of academic papers while being acceptable in the register of report writing. This problem basically requires modifications to the corpora being used. Further investigations into what kind of vocabulary and sentence patterns need to be included in a learner's writing support system for academic reports must be carried out, keeping in mind the level of the learner and the academic field. All the participants were upper intermediate level learners which points to the need to consider using a syllabus containing a graded approach in order to help learners reach a sufficient level of competency that allows them to benefit from the system. We also find problems related to difference in registers for different fields like the sciences and the humanities. Resolving these issues requires improving the content of the database and expanding it further. We intend to compile a list of adverbs that are used in reports as a step towards this. Furthermore, the results of the questionnaire carried out after the main experiment show that learners face certain problems when using Nutmeg and Natsume. The Nutmeg system should include common learner errors and give more hints and example sentences in order to make it more user-friendly. Tasks for the future include the use of error data to construct error classifiers that complement the current native language corpus-only approach. To further refine the system, we will expand and build upon the correct usage corpus data as well as the corpus of learners' writing with error tags of greater variety.

Acknowledgment

This research was funded under MEXT's Grant-in-Aid for Scientific Research (C) "Study on learner background information and learner feedback relevant within a writing support system for Japanese" (Principle investigator: Takeshi Abekawa, Research period: April, 2012 to March, 2015). We would like to present our heartfelt gratitude to all the teachers and students at the universities where we conducted the experiments. Their cooperation was crucial to this study.

References

Biber, D. & Conrad, S. (2009). *Register, genre, and style*. Cambridge University Press.

Cao, H., Yagi, Y., Kuroda, F., & Nishina, K. (2012, August). Construction of learner corpus Natane and possible application. (pp. 1–4). *5th international conference on Computer Assisted Systems for Teaching & Learning Japanese (CASTEL/J)*. Nagoya. Retrieved from http://2012castelj.kshinagawa.com/proceedings/Poster/Poster5_Cao.pdf

Halliday, M. K & Hasan, R. (1976). *Cohesion in English*. London: Longman.

Hodošček, B., Abekawa, T., Bekeš, A., & Nishina, K. (2011). Assisting co-occurrence production in report writing: Evaluation of writing assistance tool Natsume. *Journal of Technical Japanese Education, 13*, 33–40. doi:10.11448/jtje.13.33

Hodošček, B. & Nishina, K. (2011, August). On the treatment of register in writing assistance systems. (Vol. 2, pp. 522–523). *International Conference on Japanese Language Education 2011*. Tianjin, China.

Hodošček, B. & Nishina, K. (2012). Japanese learning support systems: Hinoki project report. *Acta Linguistica Asiatica, 2*(3) Lexicography of Japanese as a Second/Foreign Language (Part 2), 95–124. DOI: 10.4312/ala.2.3.95-124. Retrieved from http://revije.ff.uni-lj.si/ala/article/view/221

Ichikawa, Y. (1997). A Dictionary of Japanese Language Learners' Errors. Bonjinsha.

Imai, S. & Kuroda, F. (2012). A method of associating j-cat with other tests. Departmental Bulletin of the Tsukuba University International Student Center on Japanese Language Education, (27), 57–66.

Ishioka, T. (2008, January). Latest trends in automated essay scoring and evaluation. *Journal of Japanese Society for Artificial Intelligence (Special Issue on the Automatic Evaluation of Text), 23*(1), 17–24. Retrieved from http://ci.nii.ac.jp/naid/110006570441/en/

Nishina, K., Kamada, M., Cao, H., Utashiro, T., & Muraoka, T. (Eds.). (2012). Nihongo gakushūsien no kōchiku: Gengo kyōiku kōpasu shisutemu kaihatu [Constructing Japanese language learning: Language education, corpus and system development]. Tokyo, Japan: Bonjinsha

Teramura, H. (1990). Gaikokujin gakushūsha no nihongo goyōreishū [Collection of errors from learners of Japanese as a foreign language]. Osaka University. Retrieved from http://www.ninjal.ac.jp/teramuragoyoureishu/pdf

Yagi, Y., Hodošček, B., & Nishina, K. (2012, March). BCCWJ to gakushūsha sakubun kōpasu o riyōshita nihongo sakubun shien [Japanese writing assistance using the BCCWJ and a learner corpus]. (pp. 315-320). In Dai ikkai kōpasu nihongogaku wākushoppu yokōshū [Proceeding of the first workshop on Japanese corpus linguistics]. Dai ikkai nihongo kōpasu wākushoppu [First workshop on Japanese corpus linguistics]. Tokyo.

Yagi, Y., Hodošček, B., Abekawa, T., & Nishina, K. (2014, March). Evaluation of Error Detection in Japanese Composition Support System "Nutmeg". In *Dai gokai kōpasu nihongo wākushoppu yokōshū [Proceedings of the 5th Workshop on Japanese Corpus Linguistics]* (pp. 167–170). Dai gokai nihongo kōpasu wākushoppu [5th Workshop on Japanese Corpus Linguistics]. NINJAL.

THE LEARNER AS LEXICOGRAPHER:
USING MONOLINGUAL AND BILINGUAL CORPORA TO DEEPEN VOCABULARY KNOWLEDGE

Kristina HMELJAK SANGAWA
University of Ljubljana
kristina.hmeljak@ff.uni-lj.si

Abstract

Learning vocabulary is one of the most challenging tasks faced by learners with a non-kanji background when learning Japanese as a foreign language. However, learners are often not aware of the range of different aspects of word knowledge they need in order to successfully use Japanese. This includes not only the spoken and written form of a word and its meaning, but also morphological, grammatical, collocational, connotative and pragmatic knowledge as well as knowledge of social constraints to be observed. In this article, we present some background data on the use of dictionaries among students of Japanese at the University of Ljubljana, a selection of resources and a series of exercises developed with the following aims: a) to foster greater awareness of the different aspects of Japanese vocabulary, both from a monolingual and a contrastive perspective, b) to learn about tools and methods that can be applied in different contexts of language learning and language use, and c) to develop strategies for learning new vocabulary, reinforcing knowledge about known vocabulary, and effectively using this knowledge in receptive and productive language tasks.

Keywords: vocabulary instruction; Japanese language teaching; dictionary use; corpus use

Povzetek

Učenje besedišča je ena najbolj zahtevnih nalog pri učenju japonščine kot tujega jezika za študente, ki še ne poznajo kitajskih pismenk. Študenti se pri tem velikokrat ne zavedajo širine nabora informacij o posameznih besedah, ki je potreben za uspešno komuniciranje v japonščini. Za vsako besedo morajo namreč poznati poleg njene pisane in glasovne oblike ter pomena tudi njene oblikoslovne, skladenjske, kolokacijske, konotacijske in pragmatične lastnosti ter družbeno pogojene omejitve pri njeni rabi. V članku predstavimo rezultate ankete o rabi slovarjev med študenti japonščine na Univerzi v Ljubljani ter izbor referenčnih virov in zbirko vaj, ki smo jih sestavili z naslednjimi cilji: a) razviti zavest o različnih vidikih besedišča, tako z enojezičnega kot s kontrastivnega zornega kota; b) spoznati orodja in metode, ki se lahko uporabijo v različnih kontekstih učenja in rabe jezika; c) razviti strategije za učenje novega besedišča, utrjevanje znanega besedišča in za učinkovito rabo tega znanja v pasivnih in aktivnih jezikovnih nalogah.

Ključne besede: učenje besedišča; učenje japonščine kot tujega jezika; raba slovarjev; raba korpusov

Acta Linguistica Asiatica, 4(2), 2014.
ISSN: 2232-3317, http://revije.ff.uni-lj.si/ala/
DOI: 10.4312/ala.4.2.53-65

1. Introduction

Learning vocabulary is one of the most challenging tasks faced by learners with a non-kanji background when learning Japanese as a foreign language. Most learners are well aware of the quantitative dimension of this task from the beginning and soon develop or try to develop strategies to memorise large quantities of words and kanji characters. However, they are often not equally aware of the qualitative dimension of vocabulary and of the range of different aspects of word knowledge they need in order to successfully use Japanese. This includes not only the spoken and written form of a word and its meaning, but also morphological, grammatical, collocational, connotative and pragmatic knowledge as well as knowledge of social constraints to be observed.

In the following sections, we present some background data on the use of dictionaries among students of Japanese at the University of Ljubljana, gleaned from a small-scale questionnaire survey and from informal observations in class.

We then present a selection of resources and a series of exercises developed with the following aims: a) to foster greater awareness of the different aspects of Japanese vocabulary, both from a monolingual and a contrastive perspective, b) to learn about tools and methods that can be applied in different contexts of language learning and language use, and c) to develop strategies for learning new vocabulary, reinforcing knowledge about known vocabulary, and effectively using this knowledge in receptive and productive language tasks.

2. Background: dictionary use among Japanese language students

In order to investigate how students approach vocabulary learning, in the spring of 2013 we conducted a small-scale questionnaire on the use of reference resources (dictionaries, internet sites, mobile applications etc.) among students of Japanese in their second and third year of study at the University of Ljubljana. We found that students very often do use dictionaries and other resources during their study, but are not particularly selective when choosing which dictionary to use, and mostly look up only the most basic information.

While all of the 17 students surveyed reported that they use on-line dictionaries, only one reported having and using a dictionary in book form (Hadamitsky & Spahn 1982, a reference book of very small size) to look up the readings and stroke order of unknown characters. Six students reported having and using portable electronic dictionaries (Sharp and Casio, produced for the Japanese market) to look up Japanese-English or English-Japanese translations and unknown kanji characters.

All students reported they use on-line free dictionaries. The dictionaries and other on-line resources mentioned by students are shown in Table 1.

Table 1: On-line resources used by Japanese language students at the University of Ljubljana.

Type of resource	Resource name	no. of students (N=17)
based on WWWJDIC	jisho.org	12
	tangorin.com	4
	www.popjisyo.com	1
	rikaichan	4
other general resources	translate.google.com	6
	jaSlo	2
	reading tutor (language.tiu.ac.jp)	1
kanji resources	handwritten kanji recognition kanji.sljfaq.org	2
	kanjialive.com	1
grammar resources	j-gram (www.jgram.org)	3
	Natsume	1
	Maggiesensei.com	1

Most of the dictionaries they mentioned are based on the Japanese-English database produced by the Electronic Dictionary Research and Development Group at Monash University (EDRDG 2014), a dictionary with a very large number of lemmas but only basic information for each lemma (part of speech, translation equivalents, some stylistic labels and a crowdsourced database of usage examples). Twelve out of seventeen students used *jisho.org*, four used *tangorin.com*, four used the browser add-on *Rikaichan*, and one used *popjisyo.com*; four of the students reported using two of these, although they are only different interfaces to the same Japanese-English database.

Six students reported using the machine-translation site *translate.google.com* to look up words in several directions (Japanese to English, English to Japanese, Slovene to Japanese), while only two students mentioned other tools. Moreover, 10 out of 17 students reported they use dictionary applications on their mobile phone; some noted the name of the application (four mentioned *Kanji recognizer*, three mentioned *JED*, two mentioned *WWWJDIC* (based on EDRDG), and each of the following was mentioned by one student: *All 国語辞典*, *imiwa* (based on EDRDG), *Lexiqon* and *Aedict*), while some just answered "a dictionary on my mobile phone".

While it is clear that most students probably use freely available resources because they cannot afford expensive electronic dictionaries, it is rather surprising that none of them mentioned freely available Japanese comprehensive reference sites such as *weblio*, *yahoo jisho*, *goo jisho* or *Sanseido's Web Dictionary*. When asked about these sites in follow-up interviews, most students responded that they did know about some of these

sites, but were overwhelmed by the Japanese interface and did not feel comfortable using them.

When asked what they use these dictionaries and tools for, all students reported they look up words (translations) and unknown kanji characters, and eight out of seventeen students reported they look up how words are used in context.

Finally, when asked whether they have any difficulty with these dictionaries and tools, five students reported they do not always find the word they are looking for, three mentioned that translate.google.com is not reliable, three reported they cannot read the Japanese words they find and spend time looking up the readings, two mentioned they are not sure whether the word they find is the right one for what they mean, two reported having problems with their internet connection, and only one mentioned the problem that words and senses (translations) in the on-line dictionaries are not ordered according to frequency. These answers indicate that students tend to use dictionaries and other reference tools to look up translations of single words and character readings, and are mostly concerned with finding translation equivalents, while only some search for contexts of word use, and none apparently look up connotations, stylistic or pragmatic information.

The survey had two limitations. Firstly, the questions were open ended in order not to influence the responces, and it is therefore quite possible that students forgot to mention some of the reference sources and tools they use, or specific information that they look up less frequently. Secondly, the survey did not cover all of our students and is not representative of the whole student population. However, some tendencies were observed that are also reflected in other reports on dictionary use by learners of Japanese (Fukuda & Hiratsuka 2011, Suzuki 2012, Moroz 2013). Overall, students tend to use simple and freely available reference tools, relying on crowd-sourced Japanese-English and English-Japanese bilingual dictionaries with user- friendly interfaces. They are mostly not aware of other, more sophisticated tools and reference material, and even those students who do know about other resources mostly use them only to look up translations and readings.

On the basis of these results and of similar feedback repeatedly obtained during informal observation in class, we concluded that students need more information on other available reference sources and tools, in order to be able to select the most appropriate tool in each learning situation, and that they require coaching on the use of these tools for specific language learning needs.

3. Introducing students to different reference resources

Considering our students' reference habits, we selected a few resources that could better equip them for their learning needs and developed some exercises to help them master the use of these resources.

3.1 Dictionaries

Firstly, since many students are not aware of different freely available resources, we compiled a list of links to on-line dictionaries on the department's e-learning site[1], including

a) the Japanese-Slovene dictionary *jaSlo*[2], compiled at our department (Hmeljak Sangawa & Erjavec 2012), surprisingly not known to many of the students;

b) different interfaces to *WWWJDIC* mentioned by students themselves, including *WWWJDIC*[3] itself and its popular interfaces: *Denshi Jisho*[4], *Tangorin*[5] and *Popjisyo*[6];

c) Japanese reference sites - dictionary aggregators such as *Yahoo! dictionary*[7], *goo dictionary*[8], *kotobank*[9] and *Weblio*[10], all of which include dictionaries by major Japanese publishing houses such as Sanseido, Kodansha, Shogakukan and others;

d) the crowd-sourced Japanese-English dictionary *Eijiro*[11],

e) other sites mentioned by students, including *Google translate*[12].

This list of on-line reference sources is also presented alongside other resources (textbooks and reference books) during orientation meetings held for each Japanese language class at the beginning of the academic year, where students are encouraged to explore the resources and familiarise themselves with them.

3.2 Corpora and lexical profiling systems

The second part of the list of on-line resources includes corpora and lexical profiling systems that can be used by intermediate and advanced students to obtain more detailed information about collocational and stylistic aspects of the words they are learning.

---

1 Resource list within the departmental bulletin board "Jpn forum" at [http://e-ucenje.ff.uni-lj.si/mod/page/view.php?id=4426]

2 jaSlo [http://nl.ijs.si/jaslo/cgi/jaslo.pl]

3 WWWJDIC [http://wwwjdic.org/]

4 Denshi Jisho [http://jisho.org/]

5 tangorin [http://tangorin.com/]

6 Popjisyo [http://www.popjisyo.com/]

7 Yahoo!Japan 辞書 [http://dic.yahoo.co.jp/]

8 goo 辞書 [http://dictionary.goo.ne.jp/]

9 kotobank [http://kotobank.jp/]

10 Weblio [http://www.weblio.jp/]

11 Eijiro 英辞郎 [http://www.alc.co.jp/]

12 google translate [https://translate.google.com/]

In the last few years, quite a number of Japanese corpora and query systems have been developed, beginning with *BCCWJ*[13] developed at the NINJAL Center for Corpus Development with the concordancers *Shonagon*[14] and *Chunagon*[15] (Maekawa et al. 2014); *JpWaC*, a web corpus deployed within the lexical profiling system *Sketch Engine*[16] (Srdanović et al. 2008); its derivative *JpWaC-L* (Hmeljak Sangawa & Erjavec 2012), a corpus for learners of Japanese containing extracts from *JpWaC* ranked according to the five levels of the Japanese Language Proficiency Test specifications (JF & AIEJ 2004); the *Japanese internet corpus and query system*[17] developed at the Centre for Translation studies of the University of Leeds (Sharoff 2006); the writing support system *Natsume*[18] developed at Tokyo Institute of Technology, a lexical profiling system covering multiple monolingual Japanese corpora simultaneously (Hodošček & Nishina 2012); and the lexical profiling system *NINJAL-LWP* developed by the National Institute for Japanese Language and Linguistics and the Lago Institute of Language, applied to both the *BCCWJ*[19] (Pardeshi & Akasegawa 2011) and to the *Tsukuba web corpus*[20] (Imai et al. 2013).

Students are also encouraged to use bilingual concordances. In particular, they are introduced to two parallel concordancers. The first is the Japanese-Slovene parallel corpus *jaSlo*[21] developed at our department, a corpus of literary, academic and other web-harvested Japanese texts with Slovene translations, amounting to 760,000 Japanese tokens in 132 documents and 530,000 tokens in the corresponding Slovene translations (Hmeljak Sangawa & Erjavec 2012). The second corpus tool is *Linguee*[22], a freely available dictionary combined with a search engine that retrieves translated examples from internet-harvested bilingual texts (Calvert 2009).

4. Vocabulary exercises

After being introduced to the resources listed above, students engage in a series of vocabulary exercises and tasks, similar to those described by Frankenberg-Garcia (2012, 2014) and Montero et al. (2014). These were developed with two main aims. The first is to equip students with a knowledge of existing resources, the skills needed to use them, and strategies for choosing the most effective resource in specific situations. The second

[13] NINJAL Center for Corpus Development [http://www.ninjal.ac.jp/corpus_center/]

[14] Shonagon [http://www.kotonoha.gr.jp/shonagon/]

[15] Chunagon [http://chunagon.ninjal.ac.jp/]

[16] Sketch Engine [http://www.sketchengine.co.uk/]

[17] Leeds internet corpora [http://corpus.leeds.ac.uk/internet.html]

[18] Nastume writing support system [http://hinoki.ryu.titech.ac.jp/natsume/]

[19] NINJAL-LWP for BCCWJ (NLB) [http://nlb.ninjal.ac.jp/]

[20] NINJAL-LWP for TWC (NLT) [http://corpus.tsukuba.ac.jp/]

[21] jaSlo parallel corpus [http://nl.ijs.si/noske/jpl2.cgi/first_form]

[22] Linguee [http://www.linguee.com/english-japanese]

is to foster awareness of the different aspects of vocabulary knowledge, while encouraging autonomous learning.

We begin with simple exploratory tasks to introduce students to the use of different resources and interfaces. The first task focuses on dictionaries rather than on corpora, since dictionaries are used by all students and are already familiar to them.

4.1 Task 1: verifying the source

Since most students rely on dictionary sites or mobile applications based on the crowd-sourced Japanese-English database produced by EDRDG and are sometimes not even aware of the fact that they are looking up data from the same database using different interfaces, we prepared an exercise to raise their awareness about the structure and content of different dictionary sites, asking them to distinguish between the interface and the data source.

Students are briefly introduced to the dictionaries mentioned in 3.1 and then asked to find and compare dictionary entries for the same word in all dictionary sites.

For example, when they search for the word 留守番, they find that WWWJDIC, Denshi Jisho, Tangorin and Popjisyo provide exactly the same English translations, part of speech information and compound entry, and that even the examples (given only in WWWJDIC and Tangorin) are exactly the same, while only the amount of external links and the layout and colour of the entries are different.

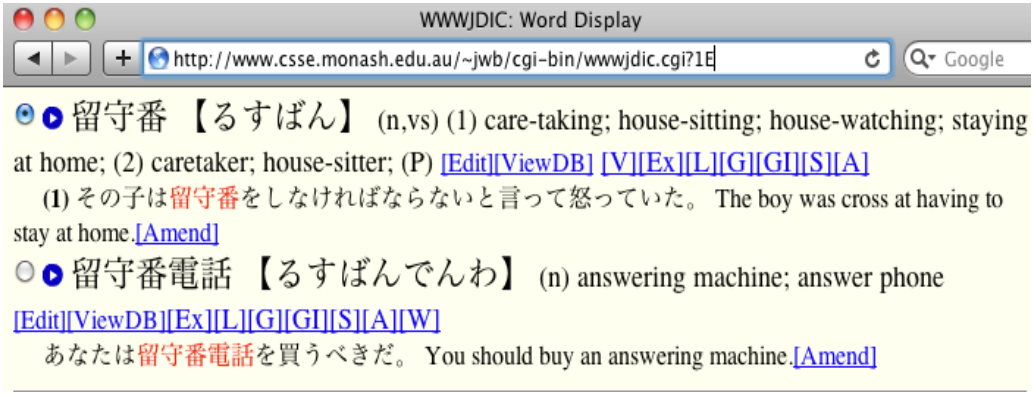

Figure 1: Results for the search string 留守番 in WWJDIC.

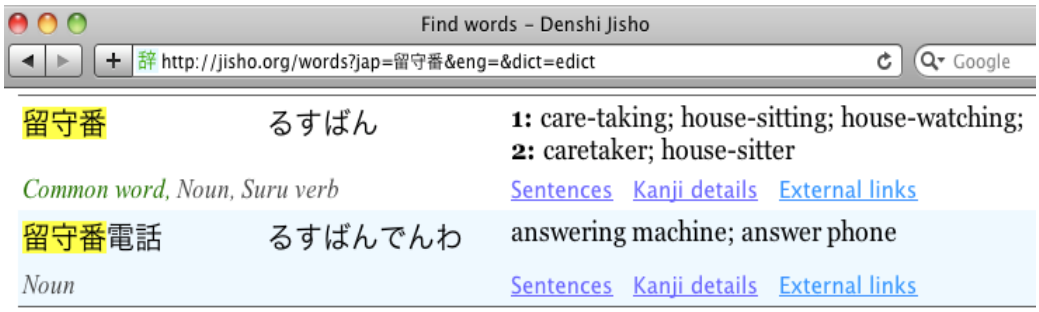

Figure 2: Results for the search string 留守番 in Denshi jisho.

Figure 3: Results for the search string 留守番 in Tangorin.

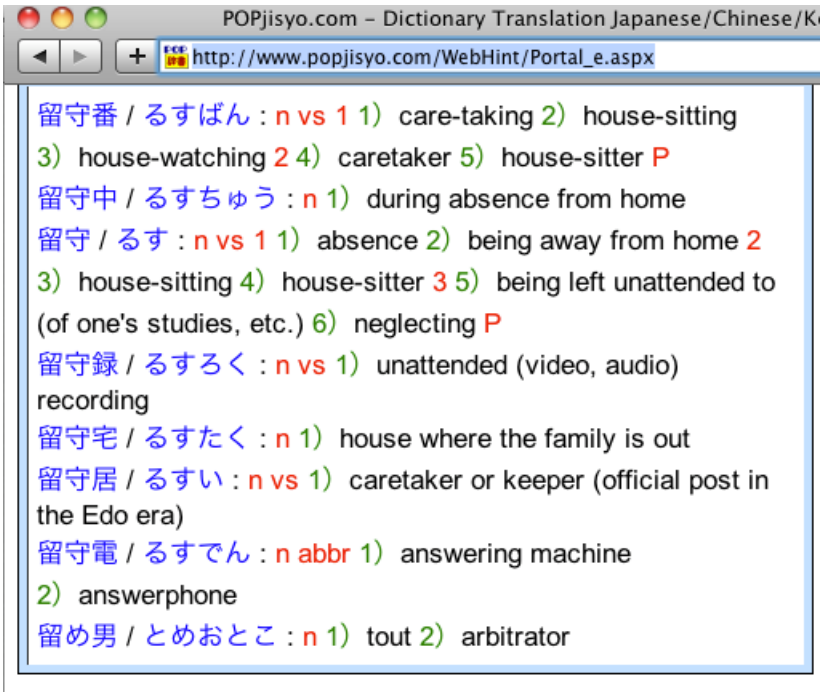

Figure 4: Results for the search string 留守番 in PopJisyo.

Further, they find that the first three of these dictionaries also provide automatically generated links to other dictionaries: *WWWJDIC* includes links to *Google search*, *Google images*, *Sanseido dictionary*, *Eijiro* on ALC, example sentences from the *Tatoeba* project, *JapanesePod101.com*, *Japanese WordNet* and *Japanese Wikipedia* (in the case of 留守番 only six of these links are available); *Denshi Jisho* offers links to *Yahoo! dictionary*, *goo dictionary* and *Google* search; and *Tangorin* offers links to *Yahoo! dictionary*, *goo dictionary*, *Eijiro*, *Weblio*, *Linguee*, *Google Translate*, *Google.com* and *Google.co.jp*.

On the other hand, they also discover that *Eijiro* offers the largest number of English translations, examples and compounds, which are different from the other dictionaries, while *weblio*, *Yahoo!dictionary*, *goo dictionary* and *kotobank* provide similar data from dictionaries that have been published also in book form by traditional dictionary publishing houses.

Finally, we point out to the students the [Edit] and [Amend] buttons in WWWJDIC and check their functioning, to make students aware that entries in this dictionary are often user generated content and sometimes need amending or editing.

With this task, students are encouraged to check the origin of the data they find in on-line reference services, to distinguish between reference data and interfaces, and to choose what works best for them considering both the ease of use and the reliability of the site.

4.2 Task 2: matching senses to corpus examples

This task is aimed at familiarising students with Japanese language corpora, their interfaces and search methods, and to help them discover the difference between dictionary descriptions or translations, which capture the main senses and uses of a word, and examples of word use in corpora, which sometimes deviate from the prototypical uses found in dictionaries. Students are presented with a monolingual dictionary definition of a word with multiple senses and its translations, and instructed to search for the same word in different corpora. Their task is to select five examples from each corpus they consult, as if they were compiling an entry for a learners' dictionary, trying to select examples that are understandable to learners at their own level, and representative of the sense described.

For example, given the following dictionary definitions and translations for the word 運動 (Shogakukan Digital Daijisen 2011, Shogakukan Progressive Japanese-English Dictionary 2012) , they matched them with examples such as those given in Table 2.

source	definition / example
Daijisen	1. 物が動くこと。物体が時間の経過とともに空間的位置を変えること。「振り子の—」「天体—」↔静止。
Progressive	1.〔物体の動き〕motion; movement
JpWaC-L	夢のない人生なんて機械の運動と変わらないじゃないか。
Daijisen	2. からだを鍛え、健康を保つために身体を動かすこと。スポーツ。「肥満防止のために—する」「—競技」
Progressive	2.〔体を動かすこと〕exercise;〔スポーツ〕sports⇒スポーツ;〔運動競技〕athletics
JpWaC-L	子供さんと遊ぶのはすごい運動ですよ。 体にも、心にも準備運動が必要なのである。 どう考えても、ものを食べて、運動もせずに痩せられるはずがない。 普通の運動靴では滑って登れない。
Daijisen	3. ある目的を達するために活動したり、各方面に働きかけること。「選挙—」「労働—」「委員になるため—する」
Progressive	3.〔働きかけ〕a drive;〔集団による〕a movement;〔特別の目的をもつ組織的活動〕a campaign
JpWaC-L	私がいた当時、すでに障害者側から反対運動が起こっていた。 ７０年代の学生運動の様子が良く分かります。 これは女性解放運動が望んだものではないであろう。

Table 2: Task 3 worksheet - matching dictionary definitions with corpus examples.

During this task students get acquainted with different corpora interfaces, practice skimming through large amounts of text, become aware of frequent compounds or collocations in which the words are used, and notice how some senses appear more frequently than others.

4.3 Task 3: comparing dictionary and corpus translations

The next task is carried out using the Japanese-Slovene parallel corpus jaSlo. Students are again given polysemous words and instructed to search for all possible translations of these words in bilingual corpus examples, in order to compile a bilingual dictionary entry for the given word. During this task they notice how some words (technical terms etc.) are mostly translated with the same equivalent, how polysemous words may have many different translations (e.g. 世話、余裕、無難、etc.), and how some words (人、また、evidential expressions, onomatopoeia etc.) are sometimes not translated at all. They furthermore explore translations (or omissions) for culturally bound terms (e.g. 就職活動, お酌 etc.), for modal expressions, adverbs (はず、せっかく、やはり、さすが、よほど), and false friends (イメージ *imidž, ドレス *dres, タレント *talent, ユニーク *unikaten, ホステス *hostesa, サービス *servis).

4.4 Task 4: translating into Japanese

The last task is again carried out using the Japanese-Slovene parallel corpus jaSlo, but in the opposite direction, searching for all possible translations (or omissions) of Slovene polysemous words, as if to compile a Slovene-Japanese dictionary entry. Distinguishing between Japanese synonyms is very challenging, and students are encouraged to first determine which word or multiword expression in the Japanese examples corresponds to the given Slovene word, and then look up these Japanese words in other corpora and dictionaries.

5. Feedback and conclusion

A portion of these exercises was tested in class, and received a mixed response. Overall, students tended to dislike too long exercises and having to browse through long concordance lists. They were frustrated and confused when they had to go through too many steps. The tasks presented above therefore need some refinement and more intermediate tasks, to gradually introduce students to different functions and search methods, with examples that are neither too difficult nor too obvious for the students' level of language competence.

Students responded positively to tasks involving adding or amending content on collaborative sites. Given the scarcity of human and financial resources for the creation of Japanese-Slovene lexicographic resources, a language combination with a very limited number of users and an even more limited number of potential bilingual

lexicographers, the involvement of Japanese language learners in a guided collaborative Japanese-Slovene dictionary project could produce useful data while at the same time raising students' awareness of the process of dictionary compilation and of the limitations of entries in existing dictionaries.

Literature

Calvert, David (2009). Tiptoeing towards TBX: strategies for terminology management at a language services provider. *Translating and the Computer 31*, 19-20 November 2009, London. 14pp. [http://www.mt-archive.info/Aslib-2009-Calvert.pdf]

Electronic Dictionary Research and Development Group (2014). *WWWJDIC: Online Japanese Dictionary Service*. [http://www.edrdg.org/cgi-bin/wwwjdic/wwwjdic?1C]

Frankenberg-Garcia, Ana (2012). Learners' Use of Corpus Examples, *International Journal of Lexicography* 25/3. 273-296.

Frankenberg-Garcia, Ana (2014). The use of corpus examples for language comprehension and production, *ReCALL* 26/. 128-146.

Fukuda, Eriko and Hiratsuka, Mari (2009). Shokyuu gakushuusha ni yoru kanjigo no imi rikai no tame no gaibu risoosu shiyou jittai chousa: denshi jisho no shiyouhou ni shouten o atete. *Hokkaidou daigaku ryuugakusei sentaa kiyou* 13/58-77. 副田恵理子、平塚真理（2009）「初級学習者による漢字語の意味理解のための外部リソース使用実態調査 : 電子辞書の使用法に焦点をあてて」『北海道大学留学生センター紀要』13/58-77. [http://hdl.handle.net/2115/45683]

Fukuda, Eriko aand Hiratsuka, Mari (2011). Nihongo rikai o shien suru gaiteki risoosu no shiyou jittai chousa: shokyuu gakushuusha no hon'yaku tsuuru no shiyou katei ni shouten o atete. *Hokkaidou daigaku ryuugakusei sentaa kiyou* 15/1-19. 副田恵理子、平塚真理（2011）「日本語理解を支援する外的リソースの使用実態調査 : 初級学習者の翻訳ツールの使用過程に焦点をあてて」『北海道大学留学生センター紀要』15/1-19.

Hmeljak Sangawa, Kristina and Tomaž Erjavec (2012). JaSlo : integration of a Japanese-Slovene bilingual dictionary with a corpus search system. *Acta linguistica asiatica*, vol. 2/3, 125-140. [http://revije.ff.uni-lj.si/ala/article/view/223]

Hodošček, Bor and Nishina, Kikuko (2012). Japanese learning support systems: Hinoki project report, *Acta Linguistica Asiatica* 2/3. 95-123 [http://revije.ff.uni-lj.si/ala/article/view/221]

Imai, Shingo, Shiro Akasegawa, and Prashant Pardeshi (2013). Tsukuba web koopasu kensaku tsuuru NLT no kaihatsu. *Dai3kai koopasu nihongogaku workshop yokoushuu.* Tokyo: NINJAL. 199-206. [http://www.ninjal.ac.jp/event/specialists/project-meeting/files/JCLWorkshop_no3_papers/JCLWorkshop_No3_26.pdf] 今井新悟、赤瀬側史朗、プラシャント・パルデシ（2013）「筑波ウェッブコーパス検索ツール NLT の開発」『第 3 回コーパス日本語学ワークショップ予稿集』国立国語研究所 199-206.

Japan Foundation, & Association of International Education Japan (2004). *Japanese Language Proficiency Test Content Specifications* (Revised ed.). Tokyo: Bonjinsha.

Maekawa, Kikuo, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den (2014). Balanced corpus of contemporary written Japanese, *Language Resources and Evaluation* 48/2. 345-371 [http://dx.doi.org/10.1007/s10579-013-9261-0]

Montero Perez, Maribel, Hans Paulussen, Lieve Macken and Piet Desmet (2014). From input to output: the potential of parallel corpora for CALL, *Language Resources and Evaluation* 48/1. 165-189.

Moroz, Ashley (2013). *App Assisted Language Learning: How Students Perceive Japanese Smartphone Apps*. Master's Thesis, University of Alberta, Department of Humanities Computing. [http://hdl.handle.net/10402/era.30241]

Pardeshi, Prashant and Shiro Akasegawa (2011). Compilation of basic verbs handbook using the BCCWJ corpus: Salient features and functions of the corpus browsing system NINJAL-LWP. *The proceedings of the symposium commemorating the completion of "the Balanced Corpus of Contemporary Written Japanese (BCCWJ)"*, 205-216. Tokyo: National Institute for Japanese Language and Linguistics. パルデシ・プラシャント、赤瀬川史朗（2011）「BCCWJ を活用した基本動詞ハンドブック作成：コーパスブラウジングシステム NINJAL-LWP の特徴と機能」『現代書き言葉均衡コーパス（BCCWJ）完成記念講演会予稿集』国立国語研究所 205-216.

Sharoff, Serge (2006). Creating general-purpose corpora using automated search engine queries. Marco Baroni and Silvia Bernardini (eds.), *WaCky! Working papers on the Web as Corpus*, Bologna: GEDIT. 63-98. [http://wackybook.sslmit.unibo.it/]

Spahn, Mark and Hadamitsky, Wolfgang (2012). Japanese Kanji & Kana: A Complete Guide to the Japanese Writing System. Clarendon, VT: Tuttle.

Srdanović-Erjavec, Irena, Tomaž Erjavec, and Adam Kilgarriff (2008). A Web Corpus and Word Sketches for Japanese, *Journal of Natural Language Processing* - 自然言語処理 15/2. 137-159.

Suzuki, Tomomi (2012). Ryuugakusei no jisho shiyou ni tsuite no jittai chousa - Toukyou gaikokugo daigaku de manabu ryuugakusei e no ankeeto chousa ano kekka to bunseki - Investigating dictionary use by foreign students. *Toukyou gaikokugo daigaku ryuugakusei nihongo kyouiku sentaa ronshuu* 38/1-21. 鈴木智美（2012）「留学生の辞書使用についての実態調査ー東京外国語大学で学ぶ留学生へのアンケート調査の結果と分析ー Investigating dictionary use by foreign students」『東京外国語大学留学生日本語教育センター論集』38/1-21 [http://hdl.handle.net/10108/70119]

YOKOHAMA PIDGIN JAPANESE REVISITED

Andrei A. AVRAM
University of Bucharest
andrei2.avram@gmail.com

Abstract

The paper is an overview of the structural features of the phonology, morphology, syntax and vocabulary of Yokohama Pidgin Japanese, an under researched contact language. The data examined are from a corpus which includes records not analyzed in previous work on this 19th century variety of pidginized Japanese.

Keywords: Yokohama Pidgin Japanese; phonology; morphology; syntax; vocabulary

Povzetek

Članek poda pregled strukturnih značilnosti fonologije, morfologije, skladnje in besedišča v Yokohama pidžin japonščini, ki je v dosedanjih raziskavah premalo zastopan sporazumevalni jezik. Podatki za raziskavo so zajeti iz korpusa, ki vsebuje dosedaj neanalizirane zapise o tej vrsti japonščine iz devetnajstega stoletja.

Ključne besede: jokohamska pidžin japonščina; fonologija; morfologija; skladnja; besedišče

1. Introduction

Yokohama Pidgin Japanese (henceforth YPJ) is a variety of pidginized Japanese, spoken in the second half of the 19th century in the multilingual setting of Yokohama[1] (Daniels, 1948, pp. 805-806; Holm, 1989, p. 593; Loveday, 1996, p. 69; Inoue, 2003; Stanlaw 2004, pp. 56-59; Stanlaw 2006, p. 181; Inoue 2004, p. 116; Inoue 2006, pp. 55-56). The YPJ speech community consisted of Japanese, Westerners (Europeans and Americans), and a sizable number of Chinese (Stanlaw 2004, p. 57; Inoue 2004, pp. 116-117; Inoue, 2006, p. 56).

As is well known, pidgins can be classified according to social criteria. Sebba (1997, pp. 26-33) proposes the following classification according to the social context of the pidgin's origins: (i) military and police pidgins; (ii) seafaring and trade pidgins; (iii)

[1] And, most probably, in two other ports, Kobe and Nagasaki.

Acta Linguistica Asiatica, 4(2), 2014.
ISSN: 2232-3317, http://revije.ff.uni-lj.si/ala/
DOI: 10.4312/ala.4.2.67-84

plantation pidgins; (iv) mine and construction pidgins; (v) immigrants' pidgins; (vi) tourist pidgins); (vii) urban contact vernaculars. YPJ emerged as a trade pidgin, therefore, it is assigned to type (ii). In terms of the social situation in which pidgins are used (Bakker, 1995, pp. 27-28), pidgins can be classified as follows: (i) maritime pidgins; (ii) trade pidgins; (iii) interethnic contact languages; (iv) work force pidgins. YPJ is a trade pidgin, i.e. it is a representative of type (ii).

YPJ is rather poorly documented in the literature on pidgin and creole languages. Previous descriptions of YPJ have looked at the sources of its lexicon (Daniels 1948) and a limited number of mainly morpho-syntactic features (Inoue 2003, 2004, 2006). Moreover, the analysis of YPJ by Inoue (2003, 2004, 2006) is exclusively based on data from a single textual source, namely Atkinson (1879).

The present paper is an attempt at providing a more comprehensive overview of YPJ. The data analyzed are from a corpus consisting of the following textual sources: a phrasebook (Atkinson 1879), a glossary (Gills 1886), a dictionary (Lentzner 1892), travel accounts (Griffis 1883, Knollys 1887), and two magazine articles (Anon.[23] 1879 and Diósy 1879). The findings are discussed with reference to other pidgins which exhibit similar characteristics.

All examples appear in the original orthography or system of transcription. The sources are mentioned between brackets. Unless otherwise specified, the translations are from the original sources. Abbreviations used: ACC = accusative; ADJ = adjective; ADV = adverb; D = Dutch; DEM = demonstrative; E = English; J = Japanese; N = noun; NEG = negator; O = object; P = Portuguese; QUANT = quantifier; S = subject; SG = singular; V = verb.

The paper is structured as follows. In section 2 I discuss the authenticity of the textual evidence. Section 3 is concerned with the phonology of YPJ. Sections 4 focuses on its morphology and syntax. In section 5 I examine the major characteristics of the vocabulary. The developmental stage attained by YPJ is discussed in section 6. The findings are summarized in section 7.

2. Authenticity of textual evidence

Given that the existence of YPJ is a matter of some debate in the literature, it is worth addressing the issue of the authenticity of the textual evidence examined.

Firstly, YPJ is mentioned by a number of contemporary authors. Diósy (1879, p. 500) refers to it as "Yokohama dialect". The same designation is used by Griffis (1883, p. 493)[3]. Other authors offer additional details. Gills (1886: 185), for instance, describes

[2] According to Daniels (1948, p. 806), the author is probably C. G. Leland.

[3] Who erroneously believes it to be "*Pigeon-English*", i.e. Pidgin English (Griffis, 1883, p. 352, n.).

it as "a species of hybrid, ungrammatical Japanese, spoken by foreigners who do not learn the language [= Japanese] accurately". Lange (1903, p. XXVIII) draws the attention of his readers to the fact that "in the ports there is a good deal of pidgin-Japanese (Yokohama dialect), which is to be avoided". Finally, Chamberlain (1904, p. 369) states that "in Japan [...] we have "Pidgin Japanese" as the *patois* in which new-comers soon learn to make known their wants to coolies and tea-house girl, and which serves as the vehicle for grave commercial transactions at the open ports". Notice that at least one author, Chamberlain (1904, p. 369), correctly recognizes the pidgin nature of YPJ.

Secondly, YPJ exhibits word-internal [ŋ], a characteristic of earlier Tokyo Japanese:

(1) *nang eye* 'long' (Atkinson,1879, p. 18) /
nangeye 'tall' (Atkinson, 1879, p. 28) (< J *nagai*)

Thirdly, the transcription of YPJ forms reflects devoicing of /i/ and /ɯ/, as well as substitution of [ʃ] for Standard Japanese [ç], both phonological characteristics of the Tokyo dialectal area:

(2) *shto* 'man' (Atkinson, 1879, p. 25), cf. Tokyo J *hito* [ʃi̥to]

Fourthly, a number of lexical items recorded in YPJ, e.g. the forms *chobber chobber* 'food, sustenance' (Atkinson 1879: 21) and *tempo* 'penny' (Atkinson 1879: 18), are attested in 19th century Japanese:

(3) *Chabu chabu komarimasŭ tempō danna san dōzo* (Griffis, 1883, p. 358)
grub be in trouble penny master please
'Please master, a penny; we are in great trouble for our grub.'

Fifthly, formally identical or similar YPJ lexical items are listed by different contemporary authors:

(4) a. *bōtō* (Diósy, 1879, p. 500) /
boto (Atkinson, 1879, p. 15) 'boat'

b. *bum-bum-funé* (Diósy, 1879, p. 501) /
boom-boom fune (Griffis, 1883, p. 30) 'man-of-war'

c. *chobber chobber* 'food, sustenance' (Atkinson, 1879, p. 21) /
chabu-chabu '(vulgar) gruel' (Diósy, 1879, p. 501)

d. *come here* (Atkinson 1879: 19),
come here (Griffis, 1883, p. 451) /
komiya (Diósy, 1879, p. 500) 'dog'

e. *dam your eye sto* (Atkinson, 1879, p. 25) /
damyuri sto (Atkinson, 1879, p. 28) /
damuraïsu h'to (Diósy, 1879, p. 500) /
dammuraisu hito (Griffis, 1883, p. 493) 'sailor'

f. *hatoba* ‘solid granite pier’ (Griffis, 1883, p. 349) /
‘a pier, or landing-place’ (Gills, 1886, p. 97) /
‘jetty’ (Knollys, 1887, p. 312)

g. *matty* ‘wait’ (Atkinson 1879) /
mate-mate ‘wait a little’ (Gills, 1883, p. 147)

Last but not least, lexical items attested in YPJ also occur – with an identical or similar form and meaning – in other Japanese-lexifier pidgins or in pidgins with Japanese as their substrate language. Consider the examples below, from Japanese Pidgin English (JPE), Thursday Island Aboriginal Japanese Pidgin (TIAJP), and Broome Pearling Lugger Pidgin (BPLP):

(5) a. YPJ *ah me* ‘rain’ (Atkinson 1897, p. 23),
TIAJP *ami* ‘rain’ (Mühlhäusler and Trew, 1996, p. 392)

b. YPJ *kooksan* ‘cook’ (Atkinson, 1879, p. 23),
TIAJP *kuk-san* ‘cook’ (Mühlhäusler and Trew, 1996, p. 392)

c. YPJ *kurrumboh* ‘gentleman of color’ (Atkinson, 1879, p. 25),
TIAJP *churumpu* ‘black man’ (Mühlhäusler and Trew, 1996, p. 392)

d. YPJ *piggy* (Atkinson, 1879, p. 21) /
peke (Diósy, p. 1879, p. 501) /
peggy (Knollys, 1887, p. 312) ‘go’,
BPLP *peke* ‘go’ (Broome Pearling Lugger Pidgin n.d.)

e. YPJ *sacky* ‘wine’ (Atkinson, 1879, p. 21),
JPE /sake/ or /sækiy/ ‘alcoholic drink’ (Goodman, 1967, pp. 51-52),
BPLP *saki* ‘grog (in general)’ (Hosokawa, 1987, p. 292),
TIAJP *sagi* ‘any drink’ (Mühlhäusler and Trew, 1996, p. 392)

Such lexical items appear, then, to have been part of the vocabulary of the pidgins in the formation of which Japanese has been involved, be it as a lexifier or as a substrate language.

3. Phonology

Given the inconsistency in the orthography or system of transcription used in the currently available sources, the phonological interpretation of the written records of YPJ can only be rather tentative in nature. Nonetheless, a number of remarks can be made with respect to the phonology of YPJ.

Consider first the deletion of the high vowels /i/ and /ɯ/ in items etymologically derived from Japanese, which reflects their devoicing in the Tokyo-Yokohama dialectal area. As is well known, when devoiced, /i/ and /ɯ/ are phonetically realized as [i̥] and [ɯ̥] respectively. Devoicing occurs in the following phonological environments: when the high vowel /i/ or /ɯ/ occurs between voiceless consonants, and in word-final position (see e.g. Author, 2005, pp. 28-33). These are also the environments in which deletion of

/i/ or /ɯ/ is attested in YPJ forms. In the examples below, deletion is indicated by the absence of the vowel letter <i> or <u> or by the use of the apostrophe:

(6) a. *arimas* 'to have' (Atkinson, 1879, p. 16) < J *arimasu* 'to be'
b. *h'to* 'person' (Diósy, 1879, p. 500) < J *hito* 'person'
c. *moots* 'six' (Atkinson, 1879, p. 18) < J *mutsu* 'six'
d. *tacksan* 'much' (Atkinson, 1879, p. 18) < J *takusan* 'much, many'
e. *watarkshee* 'I' (Atkinson, 1879, p. 15) < J *watakushi* 'I'

As already mentioned, another phonological characteristic of the Tokyo-Yokohama dialectal area reflected by YPJ forms (see also Daniels, 1948, pp. 810 and 813; Inoue, 2006, p. 58) is the substitution of [ʃ] for Standard Japanese [ç]:

(7) a. *shto* 'man' (Atkinson, 1879, p. 20) < J *hito* 'person'
b. *sheebatchey* 'stove' (Atkinson, 1879, p. 24, f.n.) < J *hibachi* 'stove'

As shown above, YPJ also has word-internal [ŋ], yet another phonological characteristic of earlier Tokyo Japanese[4]. Consider the following examples, in which [ŋ] is rendered by the digraph <ng>:

(8) a. *koongee* 'nail' (Atkinson, 1879, p. 28) < J *kugi* 'nail'
b. *tomango* 'egg' (Atkinson, 1879, p. 24) < J *tamago* 'egg'
c. *usangi uma* 'donkey' (Diósy, 1879, p. 501) < J *usag*i 'hare', *uma* 'horse'

Inoue (2006, p. 58) claims that YPJ "generally retained the CV syllabic structure of Japanese". A more accurate and complete description would be that, generally, the syllable structure of YPJ is that of Japanese, i.e. with simple syllable margins, and with /N/ as the only consonant allowed to occur in word-final codas. As in Japanese, a consequence of the rather simple nature of the syllable structure is the substantial phonological adjustment undergone by loanwords. YPJ resorts to two repair strategies for the resolution of illicit onset and codas: epenthesis and paragoge. The former is illustrated by the examples under (4), and the latter by those under (5) below:

(9) a. *bidoro* 'glass' (Diósy, 1879, p. 500) < P *vidro*
b. *buranket* 'blanket' (Diósy, 1879, p. 500) < E *blanket*
c. *sitésh'n* 'railway station' (Diósy, 1879, p. 500) < E *station*

[4] See e.g. Shibatani (1990, pp. 171-173), Author (2005, pp. 48-56).

(10) a. *bricky* ‘sheet tin’ (Diósy, 1879, p. 501) < D *blik*

b. *dontaku* ‘Sunday’ (Diósy, 1879, p. 500) < D *Zondag*

c. *madorosu* ‘sailor’ (Diósy, 1879, p. 501) < D *matroos*

The phonology of YPJ appears to have displayed considerable inter-speaker variation. This is sometimes explicitly mentioned, in the case of particular lexical items. For instance, Atkinson (1879, p. 24, f.n.) notes with respect to the word for ‘stove’ that ““Sheebatchey” is used as well as “Heebatchey””, i.e. that the use of either [ʃ] or presumably [ç] is attested. Other instances of variation are attributed to the different first languages of YPJ users. In his comments on the differences between the pronunciation of Westerners and that of the Chinese users of YPJ in the phonetic realization of [ɾ], Atkinson (1879, p. 29) writes that “foreigners as a rule rattle their “Rs” roughly, readily [...] or else ignore them altogether”, whereas a Chinese “lubricates the “R””, and provides the examples reproduced below:

(11) a. Westerner *walk-karrymasing* / *walk-kawymasing* vs.
Chinese *walk-kallimasing* ‘misunderstand (Atkinson, 1879, p. 28)

b. Westerner *am buy worry* vs.
Chinese *am buy wolly* ‘not feeling well’ (Atkinson, 1879, p. 28)

There is also variation in the form of the same YPJ lexical items recorded in different sources:

(12) a. *piggy* (Atkinson, 1879, p. 21) /
peke (Diósy, 1879, p. 501) /
peggy (Knollys, 1887, p. 312) ‘go’

b. *pumgutz* ‘punishment’ (Atkinson, 1879, p. 28) /
bonkotz ‘thrashing’ (Diósy, 1879, p. 501)

Finally, different variants are sometimes listed by the same author:

(13) a. *jiggy jiggy* / *jiki jiki* ‘make haste’ Gills (1886, p. 113)

b. *maro-maro* / *maru-maru* ‘to be somewhere’ (Diósy, 1879, p. 501)

4. Morphology and syntax

As shown by Inoue (2006, p. 60), “YPJ data do not show evidence of bound morphology”. The negator *nigh* < J *nai* is referred as a “termination” by (Atkinson, 1879, p. 17), but it is, in fact, a free morpheme. The only apparent exception is the negator *-en* < J *-en*, recorded in two YPJ forms: *arimasen* ‘not to have’ and *walk-arimasen* ‘not to understand’. However, these instances both involve high-frequency verbs and should, therefore, be regarded as unanalyzed forms.

The derivational morphology of YPJ relies on the use of several word-formation means. The most frequently used one appears to have been compounding, illustrated in the following examples:

(14) a. *chi chi amah* 'foster mother' (Atkinson, 1879, p. 25)
b. *mar gin ricky-pshaw* 'two-wheeled pony carriage' (Atkinson, 1879, p. 23)
c. *nammai kammy* 'card' (Atkinson, 1879, p. 21)
d. *niwa-tori* 'rooster' (Diósy, 1879, p. 501)
e. *yama-inu* 'wolf' (Diósy, 1879, p. 501)

In several compounds the second member is a reflex of Japanese *hito* 'person':

(15) a. *ah kye kimmono sto* 'soldier' (Atkinson, 1879, p. 25)
b. *selly shto* 'auctioneer' (Atkinson, 1879, p. 25)
c. *yakkamash' shto* 'ambassador' (Atkinson, 1879, p. 24)

Such compounds, however, occur far less often than claimed by Inoue (2006: 60), who writes that "there is one frequently observed compounding strategy with *sto/shto* 'person'". In still other compounds the second member is *mono* (< Japanese *mono* 'thing, object'):

(16) a. *ato mono* 'crupper' (Atkinson, 1879, p. 25)
b. *caberra mono* 'hat' (Atkinson, 1879, p. 15)
c. *shiroy mono* 'starch' (Atkinson, 1879, p. 24)

Suffixation is limited to the use of *-san*:

(17) a. *babysan* 'child' (Atkinson, 1879, p. 19)
b. *doctorsan* 'doctor' (Atkinson, 1879, p. 24)
c. *eejin san* 'foreigner' (Atkinson, 1879, p. 25)
d. *kooksan* 'cook' (Atkinson, 1879, p. 23)
e. *Nankinsan* 'Chinaman' (Atkinson, 1879, p. 25)

Reduplication is also found, but it is neither productive nor frequent. The available corpus contains just two instances of reduplicated forms:

(18) a. *drunky drunky* 'drunk',
cf. *drunky* 'drunk' (Atkinson, 1879, p. 28)
b. *mate-mate* 'wait a little' (Gills, 1883, p. 147),
cf. *matty* 'wait' (Atkinson, 1879, p. 20)

Furthermore, in (18a) there seems to be no demonstrable difference in meaning between the simplex and the reduplicated form. As for the other examples recorded, they are not cases of reduplication, contra Inoue (2006, p. 60), but rather "quasi-reduplicated" forms[5], i.e. not derived from a base attested in YPJ.

(19) a. *chobber chobber* 'food, sustenance' (Atkinson, 1879, p. 21)
b. *maro-maro* / *maru-maru* 'to be somewhere' (Diósy, 1879, p. 501)
c. *minner minner* 'all' (Atkinson, 1879, p. 22)
d. *para para* 'to boil' (Atkinson, 1879, p. 23)
e. *pompom* 'hammer' (Atkinson, 1979, p. 22)
f. *sick-sick* 'crank' (Atkinson, 1879, p. 20)
g. *so so* 'sew' (Atkinson, 1879, p. 21)

The lack of bound morphology and the small size of the vocabulary[6] account for the occurrence of categorial multifunctionality[7]. YPJ words can be assigned to more than one lexical category and can therefore be analyzed as lexically underspecified. Consider the examples below:

(20) a. *die job* ADJ 'strong, sound, good, able' (Atkinson, 1879, p. 19), and ADV 'well' (Atkinson, 1879, p. 23)
b. *jiggy jig* V 'to hasten' (Atkinson, 1879, p. 17), ADV 'quickly' (Atkinson, 1879, p. 17), and ADJ 'the nearest' (Atkinson, 1879, p. 19)
c. *pumgutz* V 'punish' (Atkinson, 1879, p. 22), and N 'punishment' (Atkinson, 1879, p. 28)
d. *sick-sick* N 'illness' (Atkinson, 1879, p. 17), and ADJ 'sick, ill' (Atkinson, 1879, p. 28)
e. *tacksan* QUANT 'much' (Atkinson, 1879, p. 18), and ADV 'very' (Atkinson, 1879, p. 28)

The system of pronouns and pronominal adjectives is extremely poorly developed. While personal pronouns distinguish three persons, there is no distinction in number:

(21) *watarkshee* 1SG
anatta / *anatter* and *oh my* 2SG
acheera sto 3SG

[5] As defined by Bakker (2003, p. 40): "reduplicated forms for which single forms do not exist".
[6] The recorded vocabulary of YPJ (see Daniels 1948) amounts to approximately 250 words.
[7] In the sense of Mühlhäusler (1997, p. 137).

Moreover, the only personal pronouns which are consistently used are *watarkshee* and *oh my*. The only demonstrative recorded (just once) is *kono*:

(22) *kono house*
DEM house
'this house' (Atkinson, 1879, p. 26)

Only cardinal numerals are attested.

Since there are no plural markers on nouns or pronouns plurality is inferred from the context or is expressed by e.g. numerals.

(23) *Tempo meats high kin arimas*.
penny three see be
'I see three pence.' (Atkinson, 1879, p. 18)

The Japanese case markers (particles and postpositions) do not occur. The only exception found in the corpus is case of interference with Standard Japanese:

(24) *Mado oh shimerro*. (Atkinson, 1879, p. 24)
window ACC shut
'Shut the window.'

Given the absence of case markers, possession is expressed by juxtaposition of the possessor and the possessee:

(25) *oh my tempo*
2SG penny
'your penny' (Atkinson, 1879, p. 15)

Adjectives are well represented in the corpus of YPJ. The only degree of comparison attested is the absolute superlative, formed with *num wun* preceding the adjective:

(26) *num wun your a shee*
very good
'exceptionally nice' (Atkinson, 1879, p. 25)

In the absence of an adjective, *num wan* itself has the meaning 'best' and serves to form the relative superlative:

(27) *num wun shto*
best person
'the best of men' (Atkinson, 1879, p. 20)

An overt copula *arimasu* (< Japanese *arimasu*) occurs both in equative and predicative structures:

(28) a. *Tempo arimasu*.
penny be
'This is a penny.' (Atkinson, 1879, p. 16)

b. *Kooroy arimasu*.
black be
'It is black.' (Atkinson, 1879, p. 19)

As also shown by Inoue (2006, p. 61), YPJ has no tense and aspect markers. This accounts for the fact that the temporal and aspectual interpretation relies on the context or on the use of time adverbials:

(29) a. *meonitchi [...] tacksan so so arimasu*.
tomorrow a lot sew be
'I will have plenty of work for him.' (Atkinson, 1879, p. 21)

b. *Sigh oh narrow dozo bynebai moh skosh cow*
good bye please by and by more little buy
'Good bye, please buy [in the future] some more.' (Atkinson, 1879, p. 27)

There is one invariant negator, *nigh* (< J *nai*), which occurs in post-verbal position:

(30) *Atsie sammy eel oh piggy nigh?*
hot cold colour change NEG
'Does his color change in the various seasons?' (Atkinson, 1879, p. 19)

Only a very small number of adverbs are recorded. These include:

(31) a. *bynebai* 'by and bye' (Atkinson, 1879, p. 17)

b. *coachy* 'here' (Atkinson, 1879, p. 23)

c. *meonitchi* 'tomorrow' (Atkinson, 1879, p. 21)

The following quantifiers occur in YPJ:

(32) a. *skoshe* 'a little' (Atkinson, 1879, p. 18)

b. *minner minner* 'all' (Atkinson, 1879, p. 22)

c. *tacksan* 'much' (Atkinson, 1879, p. 18)

Most question words are monomorphemic, with the exception of the alternative form for 'who' in (33e):

(33) a. *dalley* 'who' (Atkinson, 1879, p. 19)

b. *doko* 'where' (Atkinson, 1879, p. 19)

c. *ikoorah* 'how much' (Atkinson, 1879, p. 18)

d. *nanny* 'what' (Atkinson, 1879, p. 19)

e. *nanny sto* lit. 'what person', i.e. 'who' (Atkinson, 1879, p. 19)

In WH-questions question words remain *in situ*:

(34) a. *Aboorah doko?*
butter where
'Where is the butter' (Atkinson, 1879, p. 20)

b. *Mar ikoorah?*
horse how much
'How much is the horse?' (Atkinson, 1879, p. 18) [my translation]

As noted by Inoue (2006, p. 61) the word order in YPJ is SOV, as in Japanese. Generally, YPJ is typologically consistent. Consider the following parameters correlated with the SOV word order:

(35) a. possessor – possessee
oh my oh char
2SG tea
'your tea' (Atkinson, 1879, p. 15)

b. adjective – noun
die job sto (Atkinson, 1879, p. 19)
strong person
'a strong man'

c. demonstrative – noun
kono house (Atkinson, 1879, p. 26)
DEM house
'this house'

d. numeral – noun
Stoats sindoe skoshe matty.
one boatman a little wait
'Let one boatman wait.' (Atkinson, 1879, p. 20)

e. adverb – verb
coachy weedy
hither come
'come here' (Atkinson, 1879, p. 23)

There are only a few exceptions:

(36) a. *Tempo meats high kin arimas.*
penny three see be
'I see three pence.' (Atkinson, 1879, p. 18)

b. verb – adverb
Oh my piggy jiggy jig
2SG get out quickly
'Get out quickly' (Atkinson, 1879, p. 28) [my translation]

c. *Watarkshee tempo high kin nigh* ***nang eye tokey***.
1SG penny see NEG long time
'I have not seen a penny for a long time.' (Atkinson, 1879, p. 19)

Sentence coordination is achieved via parataxis:

(37) *watarkshe oki akindo, tacksan cow*
1SG big merchant a lot buy
'I am an important merchant and I buy a lot' (Atkinson, 1879, p. 26)
[my translation]

YPJ has no overt complementizers. Whether YPJ has zero complementizers is difficult to ascertain, given the scarcity and quality of the available data. Consider the following examples:

(38) a. *Start here hanash meonitchi maro maro tacksan so so arimasu*
tailor speak tomorrow pass a lot sew be
'Tell the tailor to come tomorrow and I will have plenty of work for him.' (Atkinson, 1879, p. 21)

b. *Sin turkey hanash kimmono a row.*
laundryman speak clothes wash
'Tell the laundryman to wash the clothes.' (Atkinson, 1879, p. 24)

In spite of the translation in the original, the syntax of the sentences under (38) is open to two interpretations. The clauses following the verb *hanash* 'to tell' can be analyzed either as being complement clauses or as instances of direct speech, i.e. 'Tell the tailor: Come tomorrow and I will have plenty of work for you' and 'Tell the laundryman: Wash the clothes' respectively. Compare (38) with (39):

(39) *sendo hanash drunky itchiboo sinjoe arimasen*
captain speak drunk one *bu* give be
'The captain says: I won't give the drunkard one *bu*.'
(Atkinson, 1879, p. 28) [my translation]

Since the conjunctions and conjunctive particles of Japanese have not been preserved, YPJ also relies on mere juxtaposition in other types of subordinate clauses.

Consider the examples of adverbial clauses of time (40), of reason (41), and of condition (42):

(40) *Nanny sto arimasu, watarkshee arimasen?*
who be 1SG be-NEG
'Who called when I was out?' (Atkinson, 1879, p. 19)

(41) a. *Watarkshee am buy worry oh char parra parra.*
1SG ill tea boil
'Boil me some tea because I feel ill.' (Atkinson, 1879, p. 17)
[my translation]

b. *Ginricky pshaw arimasen, mar motty koy!*
man-power carriage be-NEG horse bring
'Bring a horse because there is no man-power carriage!'
(Atkinson, 1879, p. 28)

(42) a. *Nanny sto hanash, watarkshee boto piggy.*
anyone speak 1SG boat go
'Should anyone inquire say I've gone out in the boat.'
(Atkinson, 1879, p. 21)

b. *Oh my pompom bobbery wa tarkshee pumgutz.*
2SG hammer noise 1SG punish
'If you make noise with the hammer, I'll punish you.'
(Atkinson, 1879, p. 22) [my translation]

c. *Dye die job arimasen, itchiboo sinjoe nigh.*
table good be-NEG one *bu* give NEG
'If the table is not good, I won't give you a *bu*.' (Atkinson, 1879, p. 28)
[my translation]

As can be seen, the sequencing of clauses is generally subordinate clause – main clause, as in Japanese. There is only one exception:

(43) *Nanny sto arimasu, watarkshee arimasen?*
who be 1SG be-NEG
'Who called when I was out?' (Atkinson, 1879, p. 19)

However, what appears to be an exception may well be a pragmatically-motivated so-called "afterthought" construction, which is also attested in spoken Japanese.

5. Vocabulary

Since the etyma of the lexical items found in YPJ have been discussed in great detail by Daniels (1948) this section focuses on other characteristics of its vocabulary.

Several lexical items are the outcome of reanalysis of morphemic boundaries:

(44) a. *come here* (Atkinson, 1879, p. 19) /
komiya (Diósy, 1879, p. 500) /
kumheer (Knollys, 1887, p. 311) ‘dog’
< E *come here!*

b. *dam your eye sto* (Atkinson, 1879, p. 25) /
damyuri sto (Atkinson, 1879, p. 28) /
damuraïsu h’to (Diósy, 1879, p. 500) /
dammuraisu hito (Griffis, 1883, p. 493) ‘sailor’
< E *damn your eye(s)*, J *hito* ‘man’

Also attested are lexical hybrids[8]:

(45) a. *kireen* ‘clean’ (Atkinson, p. 1879, p. 25),
cf. E *clean* and J *kirei*

b. *shiroy* ‘shirt’ (Atkinson, 1879, p. 24),
cf. E *shirt* and J *shiroy* ‘white’

Given the extremely reduced vocabulary of YPJ, synonyms would not be expected to occur. However, there are a few such instances. The synonyms are either from different source languages (43a) or from the same source language (43b):

(46) a. *am buy worry* (Atkinson, 1879, p. 17)
< J *ambai* ‘condition’, *warui* ‘bad’,
and *sick-sick* (Atkinson, p. 1879, p. 17) ‘ill’ < E *sick*

b. *die job* (Atkinson, 1879, p. 19)
< J *daijobu* ‘fine’,
and *your a shee* ‘alright’ (Atkinson, p. 1879, p. 18) < J *yoroshii* ‘good’

Two other characteristics of the YPJ vocabulary are direct consequences of the extremely small size of its vocabulary. One is the occurrence of lexical polysemy. As shown below, lexical items exhibit semantic extension, and cover a wide range of meanings:

[8] Lexical items identified across languages, given their phonetic similarity (Mühlhäusler, 1997, p. 135)

(47) a. *aboorah* 'butter, oil, kerosene, pomatum, grease'
(Atkinson, 1879, p. 20)

b. *arimasu* 'to have, to obtain, to be, to arrive, to want'
(Atkinson, 1879, p. 16)

c. *ohio* 'good morning, good day, good evening'[9]
(Atkinson, 1879, p. 17)

d. *piggy* 'to remove, take away, carry off, clear [the table], get out, remove' (Atkinson 1879: 17), 'change' (Atkinson, 1879, p. 19), 'push off' (Atkinson, p. 1879, p. 20) 'go(ne) out' (Atkinson, 1879, p. 21)

The extremely reduced vocabulary of YPJ also accounts for the use of lengthy and convoluted circumlocutions. Consider the following examples:

(48) a. *consul bobbery sto* (Atkinson, 1879, p. 25)
consul noise person
'lawyer'

b. *coots pom pom otoko* (Atkinson, 1879, p. 20)
shoe hammer man
'bootmaker'

c. *fooney high-kin serampan nigh rosoko* (Atkinson, 1879, p. 19)
ship see break NEG candle
'light house'

d. *okee abooneye pon pon* (Atkinson, 1879, p. 23)
big dangerous hammer
'earthquake'

e. *serampan funey high kin donnyson* (Atkinson, 1879, p. 25)
broken ship see master
'marine insurance surveyor'

f. *tacksan hanash bosan* (Atkinson, 1879, p. 20)
a lot speak priest
'officiating priest'

6. Developmental stage

Pidgin languages have been assigned to various developmental stages, on the basis of linguistic criteria. This well-known typology (Mühlhäusler, 1997, pp. 5-6; Siegel, 2008, pp. 2-4) distinguishes accordingly three types of pidgin: (i) pre-pidgins[10]; (ii) stable pidgins; (iii) expanded pidgins[11]. Each of these types is characterized by a specific

[9] Cf. also the comment in Anon. (1879, p. 501) on the YPJ form *ohayo*: "Foreigners use it at all hours".

[10] Also called "minimal pidgins" or "jargons".

[11] An alternative term is "extended pidgins".

set of phonological, morphological, syntactic and lexical diagnostic features (see Mühlhäusler, 1997, pp. 128-138). To these I have added one more diagnostic feature, productive morphological reduplication, since its occurrence correlates with the developmental stage of the variety at issue[12].

Consider the diagnostic features of pre-pidgins set out in Table 1 in light of the data from YPJ, discussed in sections 3, 4 and 5 ("+" = occurrence of a feature; "–" = absence of a feature):

Table 1. Diagnostic features of pre-pidgins in YPJ

Feature	**YPJ**
inter-speaker variation in phonology	+
minimal personal pronoun system	+
no copula	–
no tense and aspect markers	+
no/one adposition	+
no complementizers	+
non-productive reduplication	+
categorial multifunctionality	+
small size of vocabulary	+
reanalysis of morphemic boundaries	+
lexical hybrids	+
lexical polysemy	+
circumlocutions	+

As can be seen, with the exception of the copula, which, as shown in section 4, is found in both equative and predicative structures, YPJ exhibits features diagnostic of pre-pidgins[13].

7. Conclusions

YPJ is one of the outcomes of the language contacts which took place at the very beginning of modernity in Japan. Like other short-lived varieties emerging in similar circumstance, YPJ exhibits features typical of pre-pidgins. These characteristics have obtained via processes such as reduction, simplification, and the adoption by the groups

[12] As shown by Bakker (2003, p. 44), Bakker and Parkvall, 2005, p. 514), reduplication is not productive in pre-pidgins and stable pidgins.

[13] See also Inoue (2006, pp. 64-65), who uses a different set of criteria and reaches the conclusion that YPJ is a "restructured pidgin".

of users with different first languages of compromise solutions, with a view to attending to the immediate, bare necessities of communication.

This overview of YPJ is a contribution to a better knowledge of Japanese-lexifier pidgins, which are generally under researched. It is also hoped that the data from YPJ are relevant to the literature on pidgins and creoles, in which Japanese-lexifier varieties have figured less prominently.

References

Anon. (1879) A new dialect; or, Yokohama Pidgin. *Littell's Living Age 142 (1836)*, 496-500.

Atkinson, H. (1879) Revised and Enlarged Edition of Exercises in the Yokohama Dialect. Yokohama.

Bakker, P. (1995) Pidgins. In J. Arends, P. Muysken, and N. Smith (eds.), *Pidgins and Creoles. An Introduction*, 25-39. Amsterdam / Philadelphia: John Benjamins.

Bakker, P. (2003) The absence of reduplication in pidgins. In S. Kouwenberg (ed.), *Twice as Meaningful. Reduplication in Pidgins, Creoles and Other Contact Languages*, 37-46. London: Battlebridge.

Bakker, P. and Parkvall, M. (2005) Reduplication in pidgins and creoles. In Hurch, B. (ed.), *Studies in Reduplication*: 511-532. Berlin · New York: Mouton de Gruyter.

Broome Pearling Lugger Pidgin. n.d. en.wikipedia.org/wiki/Broome_Pearling_ Lugger_ Pidgin

Chamberlain, B. H. (1904) *Things Japanese, Being Notes on Various Subjects Connected with Japan for the Use of Travellers and Others*. London: Murray.

Daniels, F. J. (1948) The vocabulary of the Japanese ports lingo. *Bulletin of the School of Oriental and African Studies XII (3-4)*, 805-823.

Diósy, A. (1879) Japoniana curiosissima. *Littell's Living Age 142 (1836)*, 500-501.

Gills, H. A. (1886) *A Glossary of Reference on Subjects Connected with the Far East*, second edition. Hong Kong: Lane, Crawfors & Co.; Shanghai & Yokohama: Kelly & Walsh; London: Bernard Quaritch.

Goodman, J. S. (1967) The development of a dialect of English-Japanese Pidgin. *Anthropological Linguistics 9 (6)*, 43-55.

Griffis, W. E. (1883) *The Mikado's Empire*. New York: Harper & Brothers.

Holm, J. (1989) *Pidgins and Creoles*, vol. II, *Reference Survey*. Cambridge: Cambridge University Press.

Hosokawa, K. (1987) Malay talk on boat: An account of Broome Pearling Lugger Pidgin. In D. C. Laycock and W. Winter (eds.), *A World of Language: Papers Presented to Professor Stephen A. Wurm on his 65th Birthday*, 287-296. Canberra: Australian National University.

Inoue, A. (2003) Sociolinguistic history and linguistic features of Pidginized Japanese in Yokohama. Paper presented at the Annual Meeting of the Society for Pidgin and Creole Linguistics, January 2003, Atlanta.

Inoue, A. (2004) Pidginized variety of Japanese in Yokohama: Can we label it a Pidgin?. In K. Ikeda & J. Robideau (eds.), *Proceedings 2003. Selected Papers from the College-Wide Conference for Students in Languages, Linguistics, and Literature, University of Hawai'i, Mānoa*, 116-127. Mānoa: College of Languages, Linguistics and Literature, University of Hawai'i.

Inoue, A. (2006) Grammatical features of Yokohama Pidgin Japanese: Common characteristics of restricted Pidgins. In N. McGloin & J. Mori (eds.), *Japanese/Korean Linguistics* 15, 55-66.

Knollys, H. (1887) *Sketches of Life in Japan*. London: Chapman and Hall.

Lange, R. (1903) *A Text-book of Colloquial Japanese*. Tokyo: Kyobunkan.

Lentzner, K. (1892) *Dictionary of the Slang English of Australia and of Some Mixed Languages*: *With an Appendix*. Halle, Leipzig: Ehrhardt Karras.

Loveday, L. J. (1996) *Language Contact in Japan. A Socio-linguistic History*. Oxford: Clarendon Press.

Mühlhäusler, P. (1997) *Pidgin and Creole Linguistics*, expanded and revised edition. London: University of Westminster Press.

Mühlhäusler, P. and Trew, R. (1996) Japanese language in the Pacific. In S. A. Wurm, P. Mühlhäusler and D. T. Tryon (eds.), *Atlas of Languages of Intercultural Communication in the Pacific, Asia, and the Americas*, vol. II. 1, *Texts*, 373-399. Berlin · New York: Mouton de Gruyter.

Sebba, M. (1997) *Contact Languages. Pidgins and Creoles*. London: Macmillan.

Shibatani, M. (1990) *The Languages of Japan*. Cambridge: Cambridge University Press.

Stanlaw, J. (2004) *Japanese English Language and Culture Contact*. Hong Kong: Hong Kong University Press.

Stanlaw, J. (2006) Japanese and English. Borrowing and contact. In K. Bolton & B. B. Kachru (eds.), *World Englishes*, 179-200. New York: Routledge.